Emotional State Regulation in Interactive Environments:

A Psychophysiological Adaptive Approach for Affect Inductive Experiences

Pedro Alves Nogueira

Emotional State Regulation in Interactive Environments:
A Psychophysiological Adaptive Approach for Affect Inductive Experiences

Ph.D. Thesis

by

Pedro Gonçalo Ferreira Alves Nogueira

B.Sc. University of Porto 2010
M.Sc. University of Porto 2011

Main Supervisor: Eugénio da Costa Oliveira, Ph.D., University of Porto

Co-Supervisor: Rui Amaral Rodrigues, Ph.D., University of Porto

External Supervisor: Lennart E. Nacke, Ph.D., University of Ontario

December 2015

“Research is like a quantum particle with two conjugate attributes:

Where its answers lie and how fast you’re getting there.

You will never know both precisely, and at the same time.”

Dedicated to my family – not bound by blood or friendship, this would never had been quite as possible, entertaining or satisfying!

RESUMO

As emoções representam um papel significativo na forma como experienciamos e recordamos experiências passadas. Ao mesmo tempo, também representam uma das maiores limitações nos sistemas de interação pessoa-máquina modernos. Tendo isto em conta, é apenas natural que a possibilidade de capturar, interpretar e analisar emoções humanas seja um tópico de investigação popular no campo da computação afectiva. Esta tese está localizada na intersecção entre inteligência artificial, interação pessoa-máquina e psicofisiologia, reunindo conceitos emergentes de estudos de utilizador com metodologias de inteligência artificial e design de interação ou uma nova visão sobre a análise da experiência de utilizador e adaptabilidade em tempo-real baseada em afecto.

Vários estudos têm correlacionado emoções e experiência do utilizador, estabelecendo-as como um factor crucial para atingir experiências cativantes. Infelizmente, optimizar a experiência do utilizador ao adaptar ativamente a experiência afectiva proporcionada é ainda uma área sub-explorada. Isto sugere a necessidade de sistemas capazes de fornecer detecção emocional precisa em tempo-real e – principalmente – de sistemas capazes de aprender a partir das reações emocionais dos seus utilizadores de modo a poderem despertar os estados emocionais adequados de uma forma controlada, coerente e oportuna.

Implementar um tal sistema é uma tarefa caracterizada por obstáculos desafiantes. Estes obstáculos vão desde os já referidos sistemas de detecção emocional até métodos de anotação de respostas emocionais e sistemas capazes de dinamicamente alterar o seu conteúdo coerentemente. No entanto, o principal obstáculo é que não existe atualmente um sistema capaz de *1)* identificar as respostas emocionais dos seus utilizadores, *2)* criar um modelo computacional deles, e *3)* usar este modelo para dar forma à experiência afectiva. Neste sentido, muitos investigadores têm procurado apresentar métodos que possam ser usados para indiretamente moldar esta experiência afectiva. No entanto, quase todos eles assentam sobre simples mecanismos de adaptação estática que estão portanto aquém do esquema conceptual por nós proposto.

À luz destas questões, nesta tese estudamos como reações emocionais humanas a eventos de jogos digitais podem ser usadas para regular a experiência afectiva no seu geral. A metodologia proposta potencia a correlação de eventos de interação pessoa-máquina com

respostas emocionais, a partir das quais um modelo computacional de afecto pode ser extraído. Este modelo pode então ser usado para simular e prever o resultado de futuras interações, desta forma determinando o curso mais adequado para atingir a experiência afectiva desejada. Analisamos também o impacto de mecanismos de adaptação simples (estáticos) num jogo de horror procedimental, e usamos um ambiente simulado para estabelecer a capacidade dos modelos de regular a experiência afectiva dos utilizadores de acordo com um estado emocional ou padrão emocional alvos.

Adicionalmente, esta tese inclui uma *framework* conceptual para o desenho de sistemas afectivos adaptativos. Esta *framework* apresenta um desenho modular, a fim de suportar as imensas particularidades de sistemas de computação afectiva e está formalmente definido de modo a que os seus processos internos sejam facilmente replicáveis.

Muitas novas avenidas de investigação se abrem a partir daqui. A criação destes sistemas abre vastas novas possibilidades, desde agentes artificiais capazes de (um limitado) raciocínio emocional até sistemas terapêuticos, de navegação ou assistência emocionalmente sensíveis. Dum ponto de vista científico, este trabalho apresenta a possibilidade e exequibilidade de tirar proveito do poder das emoções humanas para optimizar interações pessoa-máquina, assim como um primeiro passo no sentido de sistemas capazes de uma interação emocional contextualizada e potencialmente evolucionária. Em suma, a inclusão de métricas psicofisiológicas mais detalhadas e complexas irá permitir uma análise mais compreensiva do comportamento, emoção e motivação humanos. Finalmente, a integração de novos métodos de medição e tecnologias de adaptação em aplicações de entretenimento interativo não só permitirá uma detalhada avaliação da experiência de utilizador de forma global, como também melhorar a qualidade percepcionada do conteúdo digital.

ABSTRACT

Emotions play a significant part in how we experience and remember past experiences. They also represent one of the major limitations in today's human-computer interaction systems. It is thus only natural that the possibility of interpreting and analysing human emotion has been a widely research topic within the field of affective computing. This thesis is located at the intersection of artificial intelligence, human-computer interaction and psychophysiology, bringing together emerging concepts from user research studies with methodologies from artificial intelligence and usability design; a novel take on user experience analysis and affect-driven, real-time adaption.

Various studies have correlated emotions and user experience, establishing them as a crucial factor in attaining engaging experiences. Unfortunately, improving the player's experience by actively adapting the provided affective experience is still a considerably under-explored research field. This creates the need for accurate real-time emotion detection and – primarily – to systems capable of learning from each user's reactions, so that they are able to trigger the most adequate emotional states in a logical, coherent and timely manner.

Implementing such a system is riddled with a series of challenging obstacles. These range from the aforementioned real-time emotional detection systems to emotional reaction annotation methods and systems capable of dynamically altering their content coherently. However, the main obstacle is that a system capable of *1)* identifying its users' emotional responses, *2)* creating a computational model of them, and 3) using this model to shape the affective experience, does not currently exist. In this sense, many researchers have strived to provide methods that could indirectly be used to mould the affective experience. However, most of them still rely on static, simple state adaptations and thus fall short from the grander scheme here proposed.

In light of these issues, in thesis we study how human emotional responses to digital (game) events can be used to regulate the overall affective experience. The proposed methodology enables the correlation of human-computer interaction events to emotional responses, from which a computational user affective model can be extracted. This model can then be used to simulate and predict the result of future interactions, thus determining the most adequate course to achieve the desired affective experience. We analyse both the impact of simple (static) affective adaptation mechanisms in a procedural horror game and use a simulated environment to establish the models' ability to

regulate the user's affective experience towards a specific emotional state or emotional pattern.

In addition, this thesis includes a conceptual framework for designing adaptive affective systems. This framework presents a modular design so that it can cope with the immense particularities of affective computing systems and is formally defined so that its internal processes can be easily replicated.

Many new research avenues open from here. The birth of these systems now creates vast new possibilities, ranging from artificial agents capable of emotional thought to emotionally aware therapeutic, guidance or assistance systems. From a scientific perspective, this work presents the possibility and feasibility of harnessing the power of human emotions for optimising human-machine interaction, as well as a first step towards systems capable of a contextualised and potentially evolutionary emotional interaction. In sum, the inclusion of more complex and detailed psychophysiological metrics will enable a comprehensive analysis of human behaviour, emotion, and motivation. Finally, the integration of new measurement and experience adaptation technologies in interactive entertainment applications will not only allow a detailed assessment of user's overall experience, but also improve the digital content's overall perceived quality.

ACKNOWLEDGMENTS

This doctoral thesis marks the academic coming of age of a young researcher, the culmination of knowledge on a specific topic, which could never have been realised in solitude. Were it not for the goodwill and scientific insight, motivation, support, collaboration, patience and dedication from my family, friends, colleagues, mentors, students and collaborators, this thesis would not have been possible. To each his share of the glory and the spoils of this academic battle. All of you know who you are. Thank you!

For the past four years my research has been funded by the Foundation for Science and Technology (FCT) through their doctoral scholarship programme (SFRH/BD/77688/2011), supported by the European Union. Despite this, the Artificial Intelligence and Computer Science Laboratory (LIACC) has gracefully provided virtually all of my academic dissemination and physical support -working facilities and required hardware. I am in deep gratitude for the opportunities this funding has provided me with and from the collaborations and partnerships that have emerged from it. I would also like to thank Prof. Dr. Lennart Nacke for welcoming me to his research group – the Gamer Lab – at the University of Ontario and for all his support in the most critical junctures of my research work.

In more detail, I would like to thank Prof. Dr. Eugénio Oliveira for accepting me as his doctoral student, for always securing funding, challenging me, and encouraging scientific rigor and responsibility in my work. Furthermore, thank you to my additional supervisor, Prof. Dr. Rui Rodrigues, who was always available to discuss new ideas and provide support whenever necessary.

Research is a field where getting stuck in our own ways can be oddly simple. Thus, being surrounded by a group of dedicated and interesting people is paramount to making progress at these productivity bottlenecks. During my stay at the Faculty of Engineering, I met several new colleagues – now friends – whom I would like to thank for their help, support and patience. In no particular order, thank you to: João Jacob, Hugo Machado, Nuno Cardoso, Daniel Silva, Vasco Vinhas, Lúcio Sanchez, João Almeida, Sérgio Ferreira, Inês Coimbra and António Castro. In no least degree of gratitude, I would also like to thank some older friends whom have stayed with me for the past decade or so and have graciously accepted to put up with my incessant rambling. Again in no particular order, a thank you to: José Serra, Joel Gouveia, Romeu Guimarães, and João Filgueiras. It is in no small part

due to several joint projects that I have managed to reach this far. I would like to thank my students, Hugo Fernandes, Vasco Torres, Gonçalo Silva and Rúben Aguiar, as well as my friend and colleague Luís Teófilo and Pedro Silva for their collaboration and dedication to our joint projects.

Most importantly, my greatest appreciation and gratitude goes to Vanessa Teófilo. I would like to thank you most of all for the encouragement, love, support, and, most of all, believing in me all of this time – even when I might have not. Last but not least, I want to thank my parents, Iria and Joaquim and my closest family, António, Mário, Mira and Inês. I cherish all of you, despite my sometimes-analytical outlook on life and I'm thankful for instilling in me the drive to succeed and achieve excellence. I love you for having made this possible; it's as much your success as it is mine, if not more so.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ESRB: Entertainment Software Rating Board
HCI: Human-Computer Interaction
AI: Artificial Intelligence
AC: Affective Computing
CNS: Central Nervous System
PNS: Peripheral Nervous System
PANS: Parasympathetic Nervous System
ANS: Autonomic Nervous System
SNS: Sympathetic Nervous System
GEQ: Game Experience Questionnaire
iGEQ: in-Game Experience Questionnaire
ITQ: Immersive Tendencies Questionnaire
EMG: Electromyography
EDA: Electrodermal Activity
GSR: Galvanic Skin Response
SCL: Skin Conductance Levels
EEG: Electroencephalography
UX: User eXperience
IAPS: International Affective Picture System
BP: Blood Pressure
BV: Blood Volume
HR: Heart Rate
EKG: Electrocardiogram
AIM: Affective Interaction Mechanism
BF: Biofeedback
EBF: Explicit Biofeedback
IBF: Implicit Biofeedback
ACE: Assist-Challenge-Emote
BDI: Belief-Desire-Intention
E²: Emotion Engine
PIERS: Physiological Immersion and Emotion Recognition Sub-system
GLADOS: Game Logic Alteration Daemon Operating Script
ARC: Affective Response Compendium
ARM: Affective Reaction Model
ARE²S: Affective Response Extraction and Extension Sub-system
CLEARS: Closed-Loop Emotional Adjustment and Regulation Sub-system
NN: Neural Network
MAE: Mean Absolute Error
EET: Emotion-Event Triangulation
OBO: Offline Biofeedback Optimization
ORB: Online Regulated Biofeedback

Chapter I

EMOTIONAL REGULATION

EMOTIONAL REGULATION

Recent years have witnessed the rise of systems that attempt to recognize, simulate or react to human emotions. With its roots traced to Rosalind Picard's 1995 MIT technical report (Picard, 1995), affective computing has emerged as the multidisciplinary field dealing with these issues. This drive to understand, simulate and react to human emotions on behalf of computers has long been envisioned and hypothesised by the scientific community and even considered a crucial factor for successfully completing the Turing Test. Affective computing has found many uses in real-world application fields, such as: military combat simulations, healthcare, entertainment, simulation, artificial intelligence, and education.

Affective computing is a thriving, promising field

Research in affective computing is usually tied to recognising emotions arising in human individuals, reacting to them in a specific way or simulating the processes through which emotions form and are subsequently expressed. Regarding emotion recognition, the process is most commonly implemented through the psychophysiological aspect of emotions. In other words, by measuring the physiological changes that reflect the psychological impact of external stimuli (e.g. heart rate, brain activity, muscle tension, etc.). Alternative approaches rely on less trustworthy – but also less intrusive – means, such as speech analysis and body posture. Being able to react to the user's emotional state allows us to – in theory – shape the experience itself by eliciting a desired set of emotions, thus making the experience more enjoyable and intense.

As we will see throughout this thesis, approaches towards emotionally adaptive systems rarely deviate from static, hard-coded behaviours. Matters are made worse by the often-shallow validation of presented case studies, which disperses efforts towards advancing the field's state of the art and producing real-world ready systems.

Challenges and opportunities in affective computing

In order to accelerate the aforementioned efforts, there is a need for a methodology through which dynamic emotional response profiles can be built. These profiles could then be used to endow the system with the ability to plan future interactions that would, over time, induce a more pleasurable experience. This technology would allow us to create human-adaptive software and multimedia content that would tailor themselves to the users' emotional preferences or enforce the emotional interpretation envisioned by the content developers. Although we focus on the former due to its high emotional elicitation capabilities, all of the following are immediate application candidates: digital games,

therapeutic and rehabilitation procedures, movies, music, productivity tools, guidance/assistance systems (e.g. GPS).

1.1 BACKGROUND

Computer games are a highly profitable industry

Videogames have become one of the world's favourite forms of entertainment since their creation in the early 40's by Thomas Goldsmith Jr. and Estle Ray Mann. During their evolution from simple raster graphics displayed in cathode ray tube screens to, more recently, videogame consoles and personal computers, they have become one of the most profitable industries in the world. According to the most recent data from the Entertainment Software Rating Board (ESRB), 67% of all US households play videogames, with an overall average of 8 hours a week per person (ESRB 2012). Also according to these statistics, the average gamer age is 37 years (with over 49% of the population between 18 and 49 years). This sheds a great deal of light on the industry's overall net revenues of 10.5 billion US dollars in 2009 and 18.5 billion in 2010, in the US alone (CNBC, 2010).

And also pioneers in many recent scientific breakthroughs

In the past two decades, videogames have pioneered most of the breakthroughs in a wide range of computer science fields, such as: computer animation, artificial intelligence, physics simulation and interaction techniques, among others. These achievements were propelled by a popular demand for more realistic experiences and have largely and continuously managed to produce more believable virtual consecutive improvements, we now face a period composed of small, iterative enhancements.

What makes games so engaging?

As the virtual environments on which videogames take place converge towards a photorealistic singularity, we must focus our efforts on the most promising, yet lacking areas of the experience. To do this an analysis of what drives gamers to play and constitutes a good video-gaming experience is necessary. Various studies have posed this question to comprehensive samples of the gaming community and the general consensus is that videogames are played either to: *a)* live a fantasy, or *b)* to relax from the problems of everyday life; at most times a combination of both. In either of these cases, a common trait is found: videogames must provide an engrossing experience, through which players can immerse themselves in the virtual world, as this is the ultimate goal of gaming as an activity.

Immersion first appeared in the context of early virtual reality systems and referred to how enveloped the user's senses were in the virtual world. Over the years, its use as a buzzword amongst the

general public and academia has largely contributed to an abused, vague concept. However, transversal to all the current definitions and interpretations of immersion and among other aspects, one in particular is maintained: emotions. Throughout all the definitions of immersion and player accounts of their gaming experiences, the presence of strong emotional bonds with the game world are a constant. Either through strong reactions to game events (e.g. scares, awe at the scenery, relief for overcoming a challenge) or through more lasting emotions (e.g. forming empathic bonds with the game characters or plotline), they are a sign of a good experience, investment on behalf of the player and a factor crucial in attaining the highest theoretical levels of immersion.

Immersion as more than a buzzword

While we have discussed immersion within the scope of digital games, the concept does not refer to virtual environments or videogames alone. In fact, most authors (Calleja, 2011; M Csikszentmihalyi & Rathunde, 1993; Mihaly Csikszentmihalyi, 1988; Turner, 2010) refer that one can become immersed in almost any activity, be it reading a book, riding a motorcycle or even writing a thesis. However, it is difficult to measure most of these activities outside of a laboratorial environment and other passive activities like reading a book pose little to no significant opportunities to learn from or interact with the user. Thus, videogames seem like the most logical choice. Due to their sensorial enveloping capabilities and realistic graphics, digital games envelop the human mind much more quickly and intensely than most activities, which further motivates our choice.

Why videogames as a case study?

1.2 MOTIVATION

Affective computing is gaining an ever growing community and showing practical applications, thus drawing the industry's interest in a wide range of facets, such as: gaming, healthcare, military, or educational activities. Its rapidly growing population, number of applications and target markets alone make it a field worthy of investment on research and development, especially on projects focused on applied research.

Being able to adapt any number of aspects of the presented content opens a myriad of potential applications, ranging from more human-like artificial intelligence agents to dynamic multimedia experiences or phobia / frustration management systems. Given this work's focus, the following paragraphs highlight some of the most immediate issues that may be alleviated by the work proposed in this thesis.

A common issue brought up by most consumers is that "*content lifetime*" (often referred to as "*replayability"* in videogames and

interactive play software) is steadily decreasing. This is justified by the increasing complexity of the content and the inherent production and time costs this entails. The ability for the content to become human-adaptive – i.e. altering and adapting itself continuously according to the user's own evolution over time –, could potentially alleviate this matter and provide consumers with more content for their money at the same time.

Addressable issues in digital games via affective computing

As previously referred, the main reason consumers buy videogames and general multimedia content is to immerse themselves in the world that is created by game developers. Any means of aiding this process would certainly be well accepted by the general public and add to the value of the overall experience. By doing so, it would also add to the product's value and, at the same time, serve as a selling point and competitive edge for the developing company.

Since multimedia content is almost always designed to convey a certain affective experience, content designers must always strive to convey their visions through the possible or available graphical, auditory and/or cognitive stimuli, most representative of the desired meaning. It is inevitable that by being forced to choose generic emotionally charged stimuli, some of the meanings will be lost to a percentage of the target population (e.g. a laughing clown may amuse most people, but frighten a small subset of the population). By delegating the choice of stimuli to a mechanism that is able to monitor the user's responses it becomes possible for the designer to dedicate his attention to other details and simply define high-level rules for the experience (e.g. in the following section relax the player as much as possible).

Finally, a series of research avenues become more easily studied through the collation and availability of emotional reaction models. Instead of using said models to predict human responses, they could be used to improve current representations of the human psyche and behaviour using intelligent agent software. Some potential research topics follow:

- Goal formation;
- Emotional aspects of immersion;
- Understanding and simulating human emotion formation processes;
- Automatic emotional profiling techniques;
- User experience analysis;

1.3 THESIS HYPOTHESIS

Our thesis' overarching hypothesis is divided in four parts:

I. Firstly, that even relatively simple, static gameplay adaptation mechanisms driven by players' emotional state result in (statistically) meaningful alterations in relevant user experience metrics (immersion, tension, flow, etc.).
II. Secondly, that by monitoring players' emotional states, it is possible to extract individual emotional reactions to game events in, at least, a semi-automatic fashion.
III. Thirdly, that these emotional reactions can be used to build affective reaction profiles through which players' reactions to future interactions can be estimated.
IV. Ultimately we hypothesize that using the aforementioned models, it is possible to build a system capable of regulating the user's emotional state via the presented emotional content.

Given the wide potential application of this technology, it is crucial that the focus and boundaries of our hypothesis be clearly defined. We thus restrain our conclusions to the videogames and virtual reality field. Also, while we will monitor emotional states in real-time, it is not within this thesis's main scope to develop or study novel emotion detection systems.

Scope and methodology

We will base our prototype in the existing state of the art approaches, while maintaining criticism and implementing refinements where we see fit, but will not concentrate the core of our contributions here. For the purposes of this thesis we will also consider that the means of interaction with the user are limited to a set of, although dynamic and parameterisable, well-defined and fixed stimuli set. Regarding content generation, while it is to be automatically generated within the virtual environment, we are not interested in the procedural generation of complex content such as terrain maps or architectural structures. As such, the thesis hypothesis can be summarised as follows:

"User experience in digital videogame environments can be influenced and enhanced by emotional regulation based on emotional state enforcing through a biofeedback loop. Furthermore, this loop can be implemented through the automatic evaluation of the relations between the user's emotional state data feed and an abstract stimuli set."

This biofeedback loop will be implemented through an "*Emotion Engine*" that will be able to receive a set of available stimuli and

actively regulate the player's affective experience. It will do this by measuring the player's emotional responses to the given stimuli and, over time, adapting its knowledge of his/her reactions to each stimulus. Furthermore, the system must be able to adapt to any individual or, at least, classes of individuals, and abstract the stimuli to emotional responses (e.g. vectors in Russell's circumplex model (Russel, 1980a)).

1.4 OBJECTIVES / ADDRESSED ISSUES

In this thesis work we suggest the study of how emotions can be used to regulate a user's affective experience. This work's application focus falls mainly on videogames solely based on the irrevocable fact that they pose the most comprehensive, accurate, time and cost efficient means through which to create virtual environments, while also providing the most intensive and diverse interaction mechanisms. Such a combination provides us with the highest possible degree of freedom to explore our hypotheses.

In order to study the effects of emotions on the overall user experience and further harness the user's reactions to shape the experience itself, a series of issues must first be addressed. Thus, the objectives proposed in this thesis are the following:

I. Propose a generic, conceptual architecture for the creation of emotionally adaptive systems (henceforth known as "*Emotion Engine*")

II. Define which existing state of the art methodologies are best suited for quantitatively and non-intrusively measure emotional states, while identifying limitations and potential improvements

Some objectives will necessarily be further broken down on their respective chapter

III. Develop a generic method capable of measuring the relevant emotional states. The method should provide a continuous measure of emotion in real-time, while also requiring as minimal calibration as possible. Perform a detailed comparison / validation in a relevant (to this thesis) case study with the previously identified methodologies

IV. Study the correlation between emotional states and the various facets of user experience (immersion, tension, flow, etc.), as well as the impact of static emotionally driven gameplay adaptation schemes.

V. Propose a grounded methodology to: *a)* automatically associate the emotional reactions to the eliciting interaction events, and *b)* compile the observed emotional reactions into players' affective reaction models

VI. Define an adaptation scheme to leverage the user collated data in order to reinforce a set of desired emotional states and patterns

VII. Test the created models / adaptation scheme to assess their general emotional elicitation capabilities on a wide range of emotional states.

1.5 PROPOSED FRAMEWORK

Architecture

To investigate our main research question – whether emotions can be used to regulate players' affective experience – we propose the Emotion Engine (E^2) biofeedback loop system.

In order to study the effects of emotions on user experience and using them to shape players' reactions, several issues must first be addressed:

- Develop a generic method for measuring emotional states. It should provide a continuous measure of emotion in real-time, while also requiring as minimal pre-usage calibration as possible
- Propose a methodology to: *a)* automatically associate emotional reactions to their eliciting events, and *b)* compile the user reactions into a time-evolving affective reaction profile (ARP)
- Define a method to leverage the user-collated data in order to reinforce a set of desired emotional states and patterns.

Issues posed by dynamic emotional experiences.

Each of these (sub-)issues is addressed by a separate component of our architecture. These can be seen in Figure 1:1, which presents a summary description of the architecture's general working principles. In essence, as the player experiences the game, his emotional state is reflected in his real-time physiological readings. This emotional state (ES) is then interpreted through the sensor feed by PIERS, which communicates any changes to the ARE^{2}S and CLEARS components. These components then each perform two other crucial tasks: *1)* identify and create the player's affective response profile (ARP) and *2)* regulate the player's affective experience based on what is known of his emotional preferences. ARE^{2}S creates the player's ARP by associating the changes in his emotional state output to the occurring events, which he obtains from GLaDOS – a generic game-interfacing module. On the other hand, CLEARS constantly monitors the player's ES and, if it deviates from the desired ES range, selects the most suitable implicit mechanism (game event/parameter) according to the player's ARP. The

The Emotion Engine's conceptual flow.

following sub-sections describe each of the aforementioned components in greater detail.

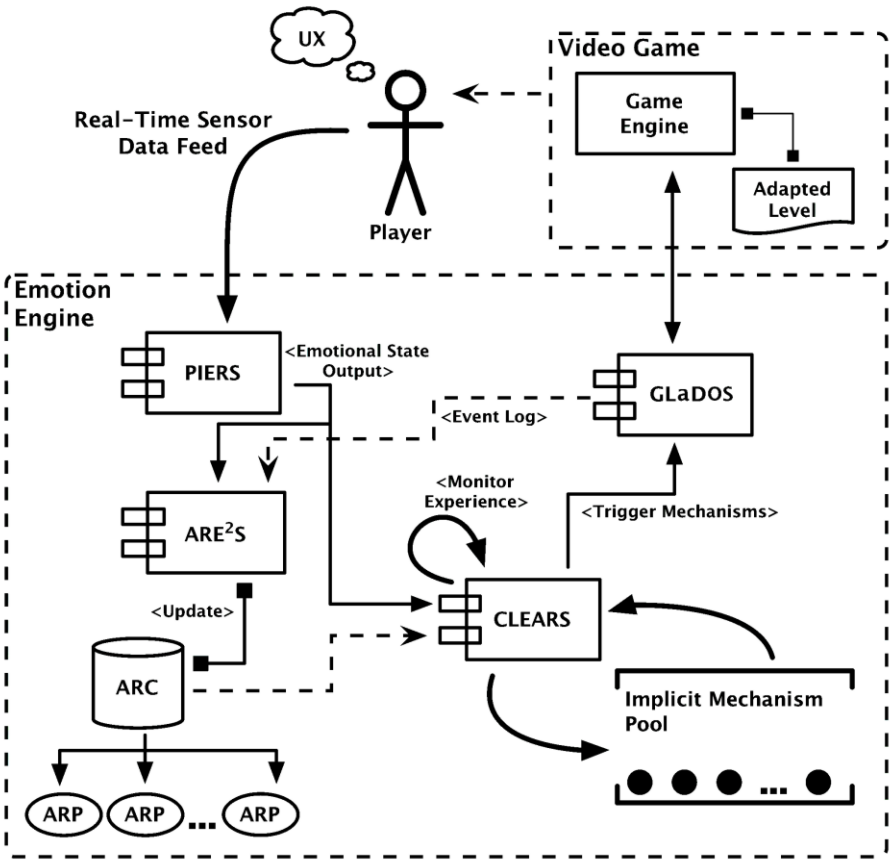


Figure 1:1: The Emotion Engine's architecture.

PIERS

The first step in the emotional regulation process is determining an, although simplified, relevant to our needs image of the user's current emotional state. To this end, we conceptualise a Physiologically-Inductive Emotion Recognition Sub-system (PIERS), which is responsible for classifying the user's emotional state in Russell's arousal/valence (AV) space (Russell 1980).

A light introduction to the emotion recognition module.

This method (as described in Chapter III) should categorize participants' AV ratings through a two-layer classification process (see Figure 1:2). The first classification layer applies several regression models to each of the four physiological inputs (SC, HR and facial EMG), which allow us to simultaneously normalize these inputs and correlate them to the AV dimensions. The second classification layer then combines the arousal and valence ratings obtained from the previous step into one final rating by using a set of rules grounded in

emotion theory literature, which combine them into one final prediction for either arousal or valence.

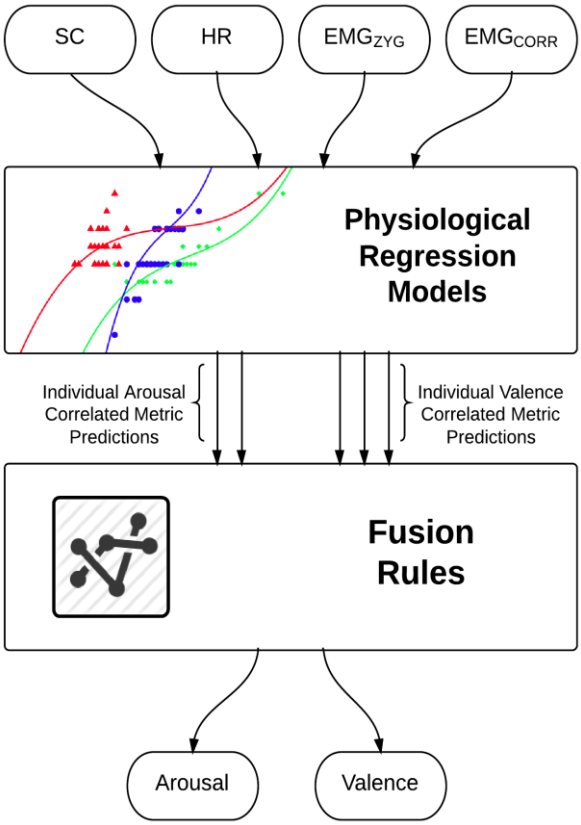


Figure 1:2. High-level overview of PIERS's architecture. Each of the physiological metrics is used to create two distinct sets of arousal and valence predictions. These predictions are then fed in parallel to a set of rules grounded in emotion theory literature, which combine them into one final prediction for either arousal or valence.

ARE²S & ARC

However precise, successfully initiating the emotion regulation process requires more than simply knowing the player's emotional state. The system must determine which regulation mechanisms should be triggered, given the player's past reactions. Extracting this information from the player/video game interaction falls upon the Affective Reaction Extraction and Extension Sub-system (ARE²S).

Extracting emotional reactions from a continuous emotional signal.

ARE²S achieves this by obtaining the game log stream (see next subsection), and linking the logged events to PIERS's emotional state output. This causality relation is determined by identifying the highest change in local maxima/minima (LMM) – herein referred to as an emotional state transition –, for each dimension of the AV space, sub-

sequent to each event. The affective reaction mapping function (Φ) is therefore formally defined as:

$$\Phi : \Lambda X \Omega \rightarrow \vec{w} \quad \sum_{i=0}^{len(\Lambda)} \overrightarrow{w_i} = 1 \ \wedge \overrightarrow{w_q} \in [0,1]$$

Where Λ is the set of possible emotional states – discretized from PIERS' continuous output – and Ω the set of possible events. Thus, the Φ function receives an emotional state λ, such that $\lambda \in \Lambda$ and an event ϖ, such that $\varpi \in \Omega$, and outputs a weight vector $\vec{w}$ that contains the probabilities of observing a transition to each of the possible states Λ ($|\Lambda| = |\vec{w}|$), if the considered event ϖ is performed at the current emotional state λ.

Learning today's reactions by forgetting yesterday's.

In order to compute the transition weights $\vec{w}_i$, we must first amass a set of observed transitions T, where we store each transition according to its chronological order. T is thus defined as a 3-tuple of the form: $t_a = (\lambda_i, \lambda_f, \varpi)$, where λ_i is the LMM emotional state preceding an event ϖ and λ_f is the LMM emotional state posterior to the same event ϖ. Updating the transition probability vector $\vec{w}$ for transition tuple t_a is computed by applying the following exponential averaging function with a learning rate α:

$$\vec{w}_i = \begin{cases} \vec{w}_i\alpha + (1-\alpha), & if\ i = f \\ \vec{w}_i\alpha, & otherwise \end{cases}$$

Each transition vector $\vec{w}$ is created with an equally distributed transition probability $p_i = |\Lambda|^{-1}$ for all possible transitions. However, these values can be manually defined in case a priori information on how users might react exists, or if we wish to set initial biases towards certain events.

By re-computing the probability values for each new observed emotional transition t_a, we expect our system will be able to learn how users react to game events over time and compensate for traditional habituation effects, leading to more accurate predictions of future reactions. Finally, the Affective Reaction Compendium (ARC) represents a database containing the full set of amassed affective reaction profiles (ARP) for each user.

CLEARS, GLaDOS & IMP

Complementary to ARE²S, the Closed-Loop Emotional Adjustment and Regulation Sub-system (CLEARS), is responsible for monitoring and eliciting the predetermined emotional states/patterns. CLEARS does so by monitoring the user's emotional state and triggering the events that

minimise the distance d to the desired ES in the AV space, such that d is given by the Euclidean distance between two emotional states (λ, λ'):

$$d(\lambda,\lambda') = \sqrt{(\lambda_a - \lambda'_a)^2 + (\lambda_v - \lambda'_v)^2}$$

Uncertainty is unavoidable when dealing with humans.

However, since we cannot be fully certain of which transition will occur (if any of the previously observed ones), we cannot base our decisions on this distance metric alone. This implies the entropy associated with the transition function must be taken into account in the event selection process. We do so using the concept of risk (Ψ), which we define as follows:

$$\Psi(\lambda,\varpi,\lambda') = 1 - \Phi(\lambda,\varpi)_{\lambda'}$$

But we can estimate how much of it there is.

In sum, the risk ψ associated with performing an event ϖ at a certain emotional state λ, while expecting a transition to a new emotional state λ' is given by the complement of the transition probability to that same state λ', obtained through Φ. The event selection process thus becomes a probabilistic optimisation problem in which we must minimise both the distance to the desired state λ_f and the risk involved in its choice – under the penalty that the occurring transition moves us even further from λ_f. We approach this issue in the following way: consider the set of all possible actions in Ω in the current emotional state λ and assume that the risk for each of the possibly occurring transitions $\psi_{\lambda,\varpi,\lambda'}$ represents the probability that an unknown and maximally entropic deviation ξ in the expected final state λ' will be observed. Furthermore, assume ξ to be a random number in the [0, ρ] interval, where ρ is equal to two times the standard deviation in $\Phi(\lambda,\varpi)$. In other words, assume that the greater the risk, the more likely it is a random transition within 95% of the distribution of previously observed transitions will occur instead. The optimal event ϖ' is the one that satisfies the following condition over all possible events in Ω:

$$\min\big(\, d(\lambda',\lambda^f) + \Psi(\lambda,\varpi,\lambda')R(0,\xi)\,\big):$$

$$\xi = 2\Gamma(d(\lambda,\lambda')[\Phi(\lambda,\varpi)])$$

Proper interfacing can make or break the experience.

Where, $R(a,b)$ denotes a function that returns a random number sampled from a uniform distribution in the $[a, b] \in \mathbb{R}$ interval, and $\Gamma(\vec{v})$ a function that computes the standard deviation of a tuple vector $\vec{v}$.

Ideally, this module should be an internal interface.

GLaDOS, the Game Layer alteration Daemon Operating Script is a generic interfacing module used to trigger the events whose implementation greatly depends on the available communication interfaces and/or game engine being used. Finally, the Implicit Mechanism Pool (IMP) consists of all the available events that can be triggered to influence the user. This static pool of events is a

configuration file with information on how each event can be triggered through GLaDOS (e.g., keystroke binds or function *callbacks*), in order to make the framework modular.

1.6 DOCUMENT STRUCTURE

This document is structured to address each of the aforementioned objectives in a systematic, linear fashion[1]. In Chapter I we have introduced the scope and motivation for this thesis. We have also introduced our conceptual framework – the "*Emotion Engine*", which will act as our blueprint for the development of all the necessary components to test our hypotheses. In light of the abundance of concepts originating outside of computer science, Chapter II provides the reader with a primer on human physiology, emotion and affective user experience. This concludes the introductory part of the thesis.

In Chapter III we analyse the current state of the art in emotional recognition methods and describe the development/validation of several of our own approaches. Chapter IV tackles the questions regarding static gameplay adaptation mechanisms, their impact on players' emotional states and user experience, as well as the correlations between the latter two.

From Chapter IV onwards, we focus on our core contribution – dynamic biofeedback techniques. Chapters V through VII describe a method for semi-automatic emotional reaction triangulation (Chapter V), the creation of affective reaction models (Chapter VI), and the model validation via statistical tests and simulated *playouts* (Chapter VII). Finally, in Chapter VIII, we present our conclusions, correlating our findings and research questions / hypotheses and outlining our contributions to each of the research areas covered throughout this thesis.

1.7 SUMMARY

In this interdisciplinary research between the fields of human-computer interaction (HCI), artificial intelligence (AI), game studies and affective computing (AC), we propose the study of how human emotions could best be used as a means for tailoring users' affective experiences in videogames. We believe emotional regulation is a powerful tool with

[1] Given the multidisciplinary nature of this thesis, we have opted for a chapter-based bibliographical style, as it helps readers to more quickly and succinctly grasp the matter at hand and clearly identify contributions.

tremendous potential that has yet to be fully explored. The main reason for this belief is that emotions are complex psychophysiological experiences of an individual, influenced by a state of mind, which arises as the result of interacting biochemical reactions and environmental interactions. As they occur at a deep and sometimes instinctual or subconscious level, emotions influence humans in a very meaningful and critical way, often overriding even rational thought. This is one of the reasons why they are main players in the formation of thoughts and, consequently, ideas. Therefore, we believe that the usage of the players' emotions as catalysts for a reactive system embedded in a game engine can be used to improve the overall user experience.

As it stands, the proposed work suggests an adaptable system that, being able to flourish in a complex environment such as an interactive virtual world has its general transferability to virtually any other area assured. If successful, this work will be applicable in any system that has a lasting and/or complex interaction scheme with its users, thus posing a significant advance in the fields of affective computing, artificial intelligence and human-machine interaction.

REFERENCES FOR CHAPTER I

Board, Entertainment Software Rating: Video Game Industry Statistics. (2012). Retrieved March 10, 2013, from 0www.esrb.org/about/video-game-industry-statistics.jsp

Calleja, G. (2011). In-Game: From Immersion to Incorporation (1st ed.). The MIT Press.

CNBC. (2010). Video Game Sales Drop 6% in 2010. Retrieved from www.cnbc.com/id/41062675/Video_Game_Sales_Drop_6_in_2010_Second_Year_of_Declines

Csikszentmihalyi, M. (1988). The flow experience and its significance for human psychology. In M. Csikszentmihalyi & I. S. Csikszentmihalyi (Eds.), Optimal experience Psychological studies of flow in consciousness (pp. 15–35). Cambridge University Press. Retrieved from http://scholar.google.com/scholar?q=Csikszentmihalyi+flow&hl=en&btnG=Search&as_sdt=1,47&as_sdtp=on#5

Csikszentmihalyi, M., & Rathunde, K. (1993). "The measurement of flow in everyday life: Towards a theory of emergent motivation" in Developmental perspectives on motivation. Nebraska symposium on motivation. (J. E. Jacobs, Ed.). Lincoln: University of Nebraska Press.

Picard, R. W. (1995). Affective Computing MIT Technical Report #321.

Russel, J. A. (1980). A Circumplex Model of Affect. Journal of Personality and Social Psychology, 39(6), 1161–1178.

Turner, P. (2010). The anatomy of engagement. In Proceedings of the 28th Annual European Conference on Cognitive Ergonomics (Vol. 44, pp. 25–27). Retrieved from http://dl.acm.org/citation.cfm?id=1962315.

Chapter II

PSYCHOPHYSIOLOGICAL CONCEPTS

A PRIMER ON HUMAN PHYSIOLOGY, EMOTION AND AFFECTIVE USER EXPERIENCE

OUTLINE

Given the multidisciplinary nature of this thesis, it is paramount that the relevant psychophysiological concepts be clarified and defined as early as possible. As such, in this chapter we introduce the reader to the human nervous system as a participant in the physiological manifestation of emotional states and emotions. We also discuss the most widely accepted operational definitions of emotions, as well as the various concepts related to immersion / user experience studies, such as presence, flow and incorporation.

2.1 HUMAN NERVOUS SYSTEM

The nervous system is the organ system responsible for coordinating the actions and sensing of humans (and animals). It communicates through a network of neurons and in most animals is divided in two parts: the Central Nervous System (CNS) and the Peripheral Nervous System (PNS). In human beings, the CNS is comprised of the brain, spinal cord and retina (although it is not included in all classifications), while the PNS consists of sensory neurons, neuron clusters called ganglia and nerves linking them together and to the CNS (Netter, 1997).

The CNS and its correlation with human physiology

The Central Nervous System is responsible for receiving information from the PNS, processing it and initiating motor control. The Peripheral Nervous System is divided into the Somatic Nervous System and Autonomic Nervous System (ANS) (Netter, 1997). The Somatic Nervous System provides the Central Nervous System with muscle or limb positioning information (proprioception) and controls skeletal muscles for body movement. The ANS (also known as visceral nervous system) is the most complicated system of the PMS and is responsible for involuntary body actions such as: glandular activity, smooth movements, and cardiac, digestive and excretory muscle control. Among these are various physiological mechanisms correlated to emotions, which justifies our study of this system. Some examples of these mechanisms are: respiratory, cardiovascular, body temperature and sudation functions (Netter, 1997).

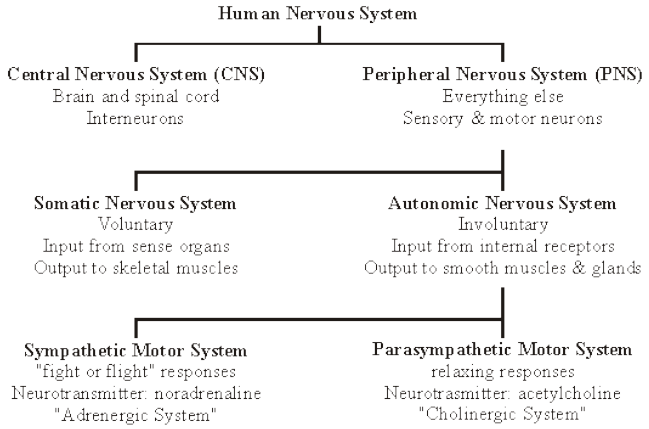


Figure 2:1: The Human Nervous System[2].

External influences on physiological activity

The ANS is subdivided into the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PANS). The enteric system, responsible for the smooth gut movements, is sometimes also considered a part of the ANS (Dodd & Role, 1997). When located in the same organ, the PNS and SNS always take control of opposite situations and functions. The SNS is usually responsible for "fight or flight" situations, such as panic attacks, physical stress, conflict or injuries. The activation of the SNS is usually accompanied by increased cardiovascular and respiratory activity, pupil dilation, a feeling of alertness and energy, as a result of energy reserve expenditure (Noback, Ruggiero, Demarest, & Strominger, 2005). On the other hand, the PNS normally controls energy conservation functions and is responsible for relaxation and digestive control. Normal PNS activity effects include heart rate decrease, intestinal activity, sexual arousal, and bladder relaxation and pupil constriction (Noback et al., 2005). A healthy subject under normal circumstances will observe a balance between the activities of both these systems. It is also noteworthy that while the SNS increases body activity, it has developed mechanisms to cope with overexerting the body's limits. These are set in place on fight or flight situations, so as to not further endanger the body (Cannon, 1929).

2.2 EMOTIONS

Emotions are complex psychophysiological phenomena that arise from human thought and its perception of the world. Given their complex, multi-faceted nature they are still a largely enigmatic process,

[2] Copyright: Mark Rothery, A level biology. URL: www.mrothery.co.uk.

representing a knowledge gap in our understanding of human behaviour. Due to the broad fields of study and communities studying them, various definitions of emotions are been suggested over the years. The Oxford Dictionary traces the origin of the term "emotion" back to the mid 16th century, denoting a public disturbance, stating the current definition originates from the early 19th century ("'Emotion,'" 2010). It further defines emotions as: "*a strong feeling deriving from one's circumstances, mood, or relationships with others*" ("'Emotion,'" 2010). Various prominent researchers in the field have proposed their own definitions of emotions. For instance, Ekman states emotions as an automatic process that is influenced by both our evolution as a species and own past experiences. This process occurs when we feel something important to our own well-being is taking place and sets a series of physiological and emotional responses to deal with the situation at hand (Ekman, 2003). Lazarus defines emotions as a complex set of reactions that have a subjective component related to the subject's mental state and an impulse to act based on a profound physiological response (Lazarus & Lazarus, 1994). Ortony, Clore and Collins, the authors of the famous OCC model, further postulate a third definition suggesting emotions are reactions to events, agents or objects that are evaluated according to the way the situation at hand is presented and interpreted (Ortony, Clore, & Collins, 1990). In fact, Kleinginna refers that there are as many as 92 different definitions of emotions in the literature (Kleinginna & Kleinginna, 1981).

Emotions as psychophysiological phenomena

Damasio, along with Spinoza, first denounced Descarte's error in believing that humans were subject of a mind-body dualism – i.e. that emotions are not merely a biological process that can be replaced by sheer logic (Damasio, 1994). Through experiences with patients suffering from localised brain lesions, Damasio found that emotions play a crucial aspect of how humans perceive, interpret and make decisions towards the world (Damasio, 1994). Thus, emotions are actually a coping mechanism through which humans prioritise the received input and deal with the overwhelming amount of variables involved in making a completely informed decision.

It is appropriate that a distinction is made between emotions and prolonged emotional states (sometimes referred in the literature as moods). An emotion is a discrete event that happens in the range of seconds to milliseconds (Stern, Ray, & Quigley, 2001); an emotional state is a prolonged, continuous and slow-changing phenomenon in which various emotions may be experienced and represents the overall emotional tendency.

Emotions and emotional states are not the same

Following the multitude of emotion definitions in the literature, Moreira et al. (Moreira, 2010) present a compilation of defining characteristics of the concept of emotion according to various authors (Ekman, 2003; Sennett, 1993; Stearns & Stearns, 1989):

Characterising emotions (partially)

- An emotion is represented as a set of feelings that are experienced, various times in a conscious manner;
- An emotional episode can have a variable duration (from seconds to milliseconds);
- Emotions set forth a multitude of physiological responses;
- Emotions reveal personal objectives;
- Emotions are experienced and are not subject to conscious choice;
- Perceiving the environment is generally a subconscious act, except in periods where such is done for long periods of time;
- There exists a fleeting moment where perceived elements are compared with past experiences, giving support to the emotional experience;
- Only after the perception process has finished, do we have conscience of feeling any emotion, after which we are permitted to reconsider it;
- There are universal emotional themes that reflect our evolutionary history. However, cultural and personal elements introduce variations in this common pattern;
- Wanting to experience or avoid a certain emotion is a great motivator in our behavioural patterns;
- Emotions provide us with mechanisms necessary for quickly responding to situations arising from the context within which we are placed;
- Emotions represent a differential between an altered state and a neutral baseline;
- Emotions are triggered by changes in our environment;
- People understand in full through interpersonal emotional analysis;

From the aforementioned definitions and characteristics three important factors are transversal and support the idea that a generic method for measuring and eliciting emotions through physiological responses can be achieved accurately across subjects:

Derived assumptions on physiological emotion recognition

- The processes through which emotions are generated are fundamentally equal to all human beings. What differs is our propensity (or bias) towards certain emotions and our interpretation of the world. In other words, it is possible to infer what emotions are felt by monitoring these physiological changes and their intensities;

- The found correlations in the literature seem to suggest equal emotional states present the same type of physiological responses across subjects – although with varying intensities due to physiological variance. Thus, it stand to reason that any method capable of accurately measuring emotional states, is therefore generalizable;
- Emotions derive from both our physiology and mental state, which in turn is determined by our interactions with the surrounding environment. This means, we can indirectly influence what emotions the subject feels;

2.3 PRESENCE

The use of the word "*presence*" derives from "*telepresence*", a term coined by Marvin Minsky in his 1980 paper entitled "*Telepresence*" (Minsky, 1980) and originally used in the field of Virtual Reality. Presence refers to the subjective experience of being physically present in a virtual environment, as opposed to being in a simulated reality, or as Slater put it, "*the feeling of being there*" (Slater & Usoh, 1994). Although initially constrained to VR systems, the use of the word "*presence*" has expanded to describe an experience realistic and believable enough to induce some degree of reality detachment on the users. Thus, its usage on other fields, such as videogame research, simulations or remote usage environments has increased. The lack of a strict formal description or applicable contexts have also contributed to its frequent misinterpretation and interchanged use with immersion, which justify the need for its clarification and distinction from our concept of immersion.

Presence is "the feeling of being there"…

… not "the act of forgetting we're actually not"…

2.4 IMMERSION

Immersion is one of the, if not the most, confusing terms in the field of user experience research. While initial concepts of immersion referred to it as the degree to which a user's sensory inputs were absorbed into the virtual environment, its use has far outgrown this limited definition. In Slater's view of immersion, the degree to which an experience is immersive is directly related solely to how faithfully the conveyed experience can be simulated (Slater, 2003). Thus, it does not matter if the user enjoys or feels about the experience itself, as long as the provided sensorial input is the same as actually living the original experience. For example, immersion is total if the sound system reproducing an orchestra is able to completely reproduce the sound in its entirety, independently of whether the listener feels like he's at the opera house or enjoys the music.

More recent definitions of immersion take into account the user's psychological state, while including the previous definition and other concepts like presence as aspects that contribute to an immersive experience. Actually, most of these views consider the psychological aspect of immersion as the major one in achieving high immersion levels and regard sensorial input fidelity as a lesser factor that simply eases the transition between the real and virtual world (Brown & Cairns, 2004; Ermi & Mäyrä, 2005; Charlene Jennett, Cox, & Cairns, 2008). Typical symptoms of an immersive experience have been documented to include:

- Heightened emotional levels (both positive and negative) (C Jennett et al., 2008);
- Temporal and spatial dissociation (Agarwal & Karahanna, 2000; Rau, Peng, & Yang, 2006; Wood, Griffiths, & Parke, 2007);
- Increased focus of attention (mirrored by real-world stimuli insensitivity and eye gaze patterns) (Cox, Cairns, Berthouze, & Jennett, 2006; C Jennett et al., 2008; Tijs, 2006);
- Increased mean power in delta and theta brainwave activity (L. E. Nacke, Stellmach, & Lindley, 2010; L. Nacke & Lindley, 2008);

Quantifying immersion

However, while the aforementioned related work has successfully established correlations between some objective measures and immersion, their application is riddled with both logistic and accuracy issues as they: *a)* are yet unable to precisely quantify how immersed a player is at each given time; *b)* can only be applied over rather large time windows (5 minutes, at least), *c)* require very intrusive (EEG) or costly (eye gaze) material, *d)* limit player's free movements and *e)* imply high setup and post-processing signal and data analysis. Despite this and while these issues can be somewhat overcome, the greatest drawback in these methods is that they base themselves in subjective user experience reports, such as Ijsselsteijn's GEQ (IJsselsteijn, Poels, & De Kort, 2008). On the other side of the spectrum, subjective experience reports – which are not without their own limitations (e.g. likely to miss out on nuanced reactions or details) –pose a more solid and tested approach that is easily applicable and comparable within a large player population.

Full immersion requires firstly feeling there and then, forgetting we're actually not.

In these takes on immersion, emotions are taken as important (albeit yet underexplored) aspects of immersive experiences. Most authors refer emotions as a means through which players form empathic relations with the virtual world or characters, thus increasing

immersion (Baños et al., 2004; Douglas, 2000; Ermi & Mäyrä, 2005). For example, Cairns describes *"empathy with the (virtual) characters and atmosphere"* as the main barriers to be overcome in achieving the final stages of immersion (Brown & Cairns, 2004) (i.e. the highest the empathy and atmosphere, the easier it is to transverse the barrier towards full immersion). As one of our main objectives is to study the effects of emotions on subjective user experience concepts such as immersion and the complex relation between immersion and emotions, we further discuss this topic in sub-section 2:5.

In sum, immersion should be viewed as *a fundamentally psychological state that is influenced by the user's own subjective – often emotional – interpretation of the experience and the believability of the provided sensorial input*.

2.5 IMMERSION AND EMOTIONS

In the previous section we discussed the various operational definitions of immersion portrayed in the literature and how it has been measured thus far. Throughout our analysis, a clear trend to associate emotions with the immersive process was visible, which should come as no surprise since it is, as we have seen, a profoundly psychological phenomenon. In this section we will examine exactly how emotions have been known to influence immersion and conjecture on how they can be harnessed to deepen it.

Components of immersion

In their layered view of immersion, Cairns et al. describe emotions and empathy towards the game world, characters and plotline as a major factor in immersion (Brown & Cairns, 2004). Seeing as they view immersion as a gradual process (which we have established it is), they argue that emotion only comes into play when the subject has been given enough time[3] to familiarise himself with the game world and controls. However, in their work, emotion and empathy are the main barriers to attain the highest level of immersion (Charlene Jennett et al., 2008). Ermi and Mäyrä also support the idea that emotion is a relevant factor in immersion. In their work, they propose that immersion has three key dimensions – sensory, challenge and imaginative immersion (Figure 2:2) – and place emotions at the centre of the latter (Ermi & Mäyrä, 2005).

[3] Exactly how much time it takes for the player to familiarise himself with the game varies but a player accustomed to the genre takes less than 15 minutes to surpass these barriers. For example, it is not uncommon for gamers to exhibit high levels of emotional responses when playing survival horror games (e.g. Dead Space (Schofield et al., 2008), Silent Hill (Nakazawa et al., 2003)).

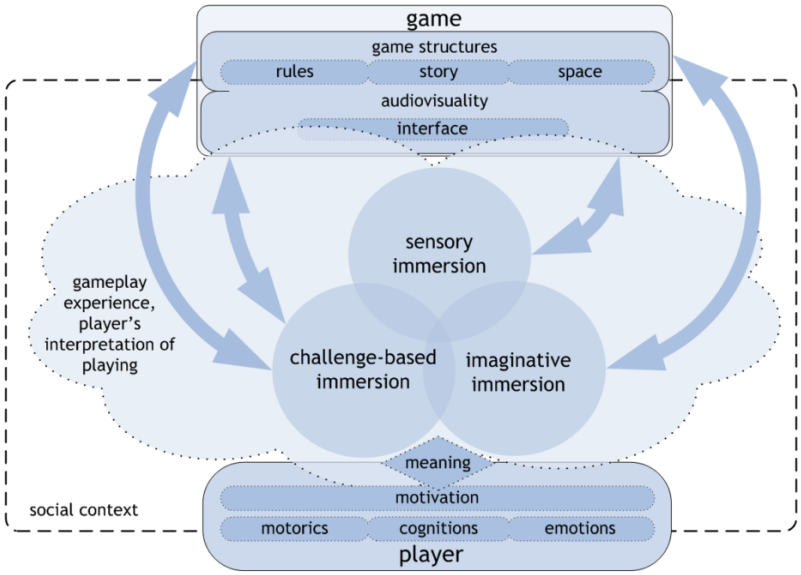


Figure 2:2: The SCI-model (sensory, challenge, imaginative), which "*identifies the three key dimensions of immersion that are related to several other fundamental components, that have a role in the formation of the gameplay experience*" (Ermi & Mäyrä, 2005).

Emotions as an immersion factor across game genres

Ermi and Mäyrä then proceed to analyse the user-reported levels of each immersion dimension in a series of popular videogames, concluding that imaginative was more present in games with enticing characters and plotlines. A conclusion, once again, in line with Cairn's last barrier of immersion; empathy and emotions. The authors also conclude, as can be seen by inspecting Figure 2:3 that imaginative immersion is the most rare and difficult one to achieve.

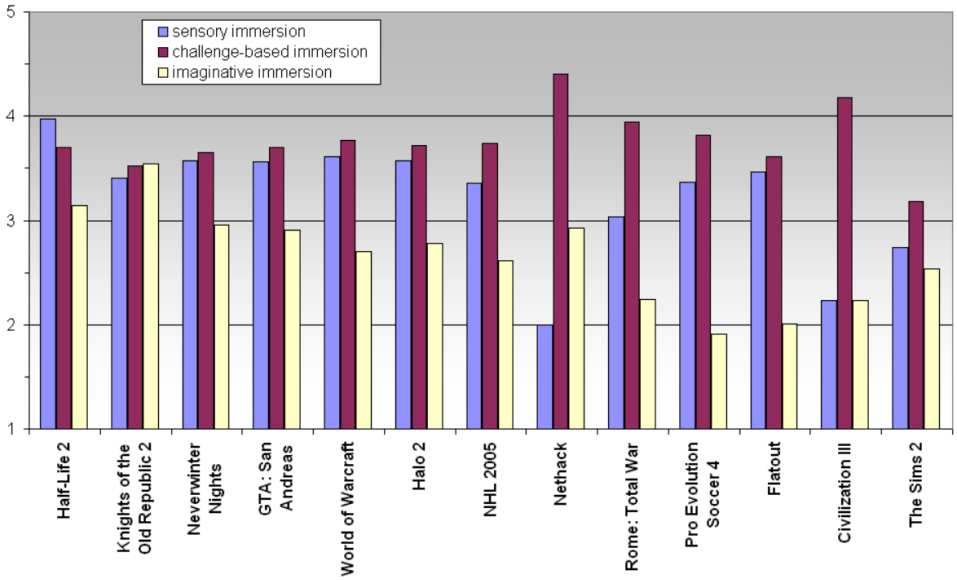


Figure 2:3: Average amount of each immersion dimension reported through questionnaires by players in various popular digital games. Games on the left hand side presented the highest amount of overall immersion.

Coincidentally, Baños et al. almost simultaneously proved these hypotheses by experimentally verifying that emotions significantly impacted the user's sense of presence (immersion as transportation) (Baños et al., 2004). Their studies showed that while the employed physical apparatus (computer monitor, HMD or wall projector) influenced presence, emotions helped the subjects attain even higher presence levels, as well as in dissolving the medium. Later work on this topic by Riva et al. further reinforces these claims as they observed: "*a circular interaction between presence and emotions… the feeling of presence was greater in the "emotional" environments… (and) the emotional state was influenced by the level of presence*" (Giuseppe Riva et al., 2007).

While these findings support Cairn's work, they also posit an interesting question: "*If emotions help in dissolving the medium (i.e. break the first two barriers of immersion: controls and environment familiarisation), do they also come into play earlier than the second immersive stage, as proposed by Cairns?"*. If so, immersion cannot be seen as such a linear process and further justifies our choice of the incorporation metaphor, where all aspects interplay and occur parallel to each other.

In fact, the work presented by Douglas (Douglas, 2000) and Jennett (Charlene Jennett et al., 2008) seems to support the hypothesis that different aspects interplay with each other in significant manners. Douglas et al. present compelling arguments in favour of positive emotions arising from complex narratives, while Jennett et al. refers that emotions partially stem from empathic bonds with characters, obtained through plot involvement (narrative involvement).

With our hypothesis that videogames (and multimedia experiences in general) generate emotional responses, we now turn to a more complex aspect of this phenomenon: "*Which emotions are generated, and how?*"

Do some emotions enhance immersion more heavily or it is dependent on the game genre?

Ravaja et al. have provided some initial answers to this question in his 2004 NordiCHI paper (Ravaja et al., 2004). Through a series of tests with 34 undergraduate students, the authors found that games produce emotional patterns and each affects a limited scope of emotions (which we will from now on refer to as the game's *emotional spectrum*). These patterns are intrinsically tied to the gameplay mechanics, while the emotional spectrum is connected to the game's setting/genre. However, one point where the paper falls short is by not exploring any of the possible correlations between the reported immersion levels and the

elicited emotional patterns/spectrums. The existence of such correlations is a key hypothesis of our thesis and, as such, we will conduct experiments to determine its veracity.

Upon contemplating the possible synergies between emotions and immersion, a pertinent question arises: "*If immersion entails more satisfying experiences and emotions can influence immersion, how can we capitalise on this correlation?*" To answer this question, we should first turn to why gamers play videogames and then to the main barriers to emotional content communication.

Lazzaro's 2005 paper "Why We Play Games" provides some insight as to why people indulge in videogames as a leisure activity. He states that:

"*Adults in this study, enjoy filling their heads with thoughts and emotions unrelated to work or school… They value the sensations from doing new things such as dirt-bike racing or flying, that they otherwise lack the skills, resources, or social permission to do. A few like to escape the real world; others enjoy escaping its social norms. Nearly all enjoy the feeling of challenge and complete absorption. The exciting and relaxing effects of games is very appealing and some apply its therapeutic benefits to get perspective, calm down after a hard day, or build self-esteem*" (Lazzaro, 2005).

These statements hint at the impression that people play videogames either to escape from reality or to feel some sort of emotional stimuli, be it what may. This pursuit of emotional arousal is a constantly referred factor in the literature where user statements are cited (Gow, Cairns, Colton, Miller, & Baumgarten, 2010), (Brown & Cairns, 2004; Charlene Jennett et al., 2008), and what is referred in media psychology as *excitatory homeostasis* (Vorderer & Bryant, 2006). From Lazzaro's quote we also draw an important conclusion regarding the emotional stimuli sought by gamers. Their search may be either for immediate adrenaline rushes, or relaxing experiences. This leads us to the conclusion that, like a film, while each individual game only caters to a certain emotional spectrum, videogames in general are able to approach a wide range of emotions. This has been partially corroborated by Jennett when she verified both positive and negative emotions run high in gaming experiences (C Jennett et al., 2008).

Game designers have come up with a series of techniques to make the experience as involving and emotionally charged as possible by using:

- Emotionally evocative graphics and environments (e.g. Sony's Ico (Sony, 2001) breathtaking architectural and lighting style or Naughty Dog's Uncharted series (Naughty Dog, 2009) exotic, abandoned ruins: see Figure 2:4);
- Fast-paced gameplay mechanics;
- Scare tactics (i.e. building up tension up to a confrontation with a frightening enemy);
- Context-sensitive events and story narration;
- Scripted gameplay (i.e. forcing the player to go through dramatic events in a controlled environment);
- Event or enemy randomization to escape from predictable design patterns in high content games (e.g. the Bethesda's Elder Scrolls (Bethesda, 2010) series);

How to elicit emotions in digital games?

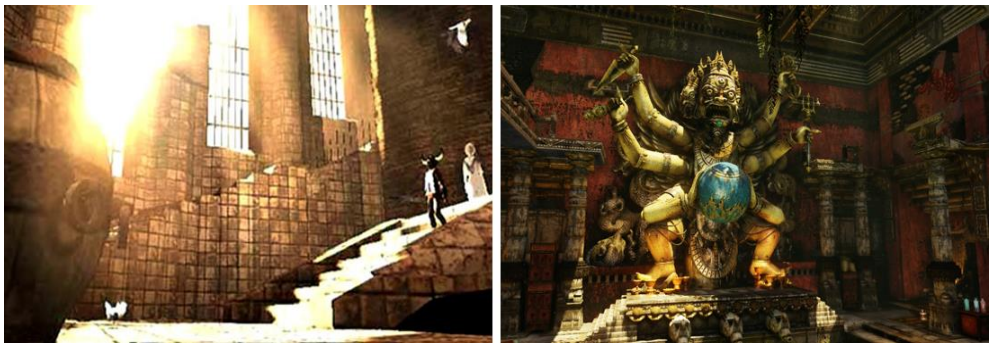


Figure 2:4: Evocative game level design and architectures. From left to right: Gameplay screenshots from ICO and Uncharted 2: Among Thieves.

How to elicit them reliably and with a minimal set of assumptions?

However, while most of these techniques work when duly applied, it is impossible to cater the experience to every player, as some will be acquainted with the genre, company's design rules or simply lose interest more quickly. Personal misinterpretation is an equally intractable issue, as it is not possible to know beforehand each player's reaction to any possible event. Extending current usability and user experience testing to account for emotional reactions can potentiate the creation of emotional design guidelines. Further expanding on this concept – and re-iterating our hypothesis – we propose that being able to monitor players' reactions in real-time would allow us to build individual emotional reaction profiles to game events. With these, it would be possible to dynamically tailor future events to modulate the user's emotional response in such a way that immersion levels (or any other aspect of the user experience as a whole) would be optimal.

2.6 FLOW

Initially proposed in 1988 by Mihaly Csikszentmihalyi, flow is a concept that refers to a feeling of optimal experience, where there is a perfect balance between the user's abilities and the challenges presented to him

(Mihaly Csikszentmihalyi, 1988). Csikszentmihalyi describes flow as "*a state in which people are so involved in an activity that nothing else seems to matter*" (Csíkszentmihályi, 2008). Flow is often accompanied by experiencing high positive emotional levels, creativity and temporal-spatial disconnections with the real world, often neglecting basic physiological needs, such as sleeping or eating. Csikszentmihalyi describes nine factors that commonly accompany a flow experience (M Csikszentmihalyi & Rathunde, 1993):

- Clear goals (expectations and rules are discernible; goals are achievable and aligned with one's abilities and skills);
- Direct and immediate feedback providing a clear knowledge of one's current situation within the experience;
- Goals and required skill sets are both equally matched and high;
- Concentration is limited to a narrow field of attention. Action and awareness are merged (i.e. one is always focused on their actions);
- Loss of self-consciousness (spatial dissociation) and insensibility to negative feelings;
- Inability to think about failing. One's complete focus is on the task at hand, nothing more;
- Dissolution of the ego (no self-consciousness about one's performance);
- Distorted sense of time (temporal dissociation);
- The experience is "autotelic" (i.e. done for its own sake);

While the concept of flow does evidently share some characteristics with immersion, it also undoubtedly diverges in some key aspects. While, spatial and temporal distortion, along with loss of self-consciousness and high concentration levels are shared traits with immersion, the inability to think about failing or feel negative feelings are not. One could argue that clear goals, immediate feedback and an autotelic experience are important aspects of immersive experiences, but this does not detract from the fact that immersive experiences are not intrinsically emotionally positive. For instance, in the Silent Hill (Nakazawa, Yamaoka, Ito, & Owaku, 2003) videogame series (Figure 2:5), the player is forced to wander alone through a deserted small town riddled with disturbing creatures and psychological terror attacks. Indubitably, high positive emotions have no place in such an experience. However, the series has been largely acclaimed as one of the most immersive experiences in the gaming world, along with other survival horror games such as BioShock (Hellquist, Levine, & Schyman, 2007), Dead Space (Schofield et al., 2008) or Fatal Frame (Kikuchi, 2012) (Figure 2:5). In fact, Jennet et al. have corroborated the thesis that

Flow is often seen as a full immersion state characterised by positive emotions.

negative emotional levels run high in immersive experiences (C Jennett et al., 2008). This leads to the conclusion that while flow does share traits with immersion, they are not the same, although it is possible that flow describes an immersive experience, albeit a highly positive one.

Figure 2:5: Various highly immersive horror games. Left to right, top to bottom: Dead Space 2 (Schofield et al., 2008), Fatal Frame (Kikuchi, 2012), BioShock (Hellquist et al., 2007) and Silent Hill 3 (Nakazawa et al., 2003).

2.8 EMOTIONAL REGULATION

Emotion regulation is considered an important driver of affective computing (Picard, 1995). However, it is a concept closely tied to behavioural psychology and it requires a definition in a computational context. According to Gross and Thompson (J. J. Gross & Thompson, 2009), emotion regulation is normally used to refer to "*an individual's ability to evaluate, understand and modulate their emotional responses to events according to some strategy*". These strategies may target managing uncomfortable emotions, engaging in socially acceptable conduct or avoiding manic or depressive behaviour. While emotional regulation may be self-induced, other uses of the term may refer to employing a certain type of therapy, through which emotions are used to influence the patient in some way (J. J. Gross & Thompson, 2009).

In our work, we redefine the term within the context of video games and interactive multimedia scenarios as: *A process through which users' emotional responses are induced by a set of carefully chosen stimuli, aimed at eliciting a certain emotional state or pattern over time*. The process is further delineated by three separate phases that occur

continuously and in parallel:

I. Modelling and monitoring the user's emotional state in real-time
II. Providing emotionally-charged stimuli as a means through which the emotional state is modulated to the desired states or patterns
III. An observation of the user's emotional responses to the stimuli set, while taking them into context and storing them for future reference.

REFERENCES FOR CHAPTER II

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. MIS Quarterly, 24(4), 665–694.

Baños, R. M., Botella, C., Alcañiz, M., Liaño, V., Guerrero, B., & Rey, B. (2004). Immersion and emotion: their impact on the sense of presence. Cyberpsychology & Behavior : The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society, 7(6), 734–41. doi:10.1089/cpb.2004.7.734

Brown, E., & Cairns, P. (2004). A grounded investigation of game immersion. In Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04 (p. 1297). New York, New York, USA: ACM Press. doi:10.1145/985921.986048

Cannon, W. B. (1929). Bodily changes in pain, hunger, fear, and rage. New York: Appleton-Century-Crofts.

Cox, A. L., Cairns, P., Berthouze, N., & Jennett, C. (2006). The Use of Eyetracking for Measuring Immersion. workshop on What have eye movements told us so far, and what is next? In Twenty-Eighth Annual Meeting of the Cognitive Science Society (CogSci2006) (p. N/A). Vancouver, Canada.

Csikszentmihalyi, M., & Rathunde, K. (1993). "The measurement of flow in everyday life: Towards a theory of emergent motivation" in Developmental perspectives on motivation. Nebraska symposium on motivation. (J. E. Jacobs, Ed.). Lincoln: University of Nebraska Press.

Damasio, A. (1994). Descartes' Error: Emotion, Reason, and the Human Brain (p. 336). Penguin Books.

Dodd, J., & Role, L. W. (1997). The autonomic nervous system. Principles of neural science. (E. R. Kandel, J. H. Schwartz, & T. M. Jessel, Eds.) (3rd ed.). Appleton & Lange.

Douglas, Y. (2000). The pleasure principle: immersion, engagement, flow. In Proceedings of the eleventh ACM on Hypertext and hypermedia (pp. 153–160).

Ekman, P. (2003). Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life (1st ed.). Times Books.

Emotion. (2010). In Oxford Dictionaries. Oxford University Press.

Ermi, L., & Mäyrä, F. (2005). Fundamental components of the gameplay experience: Analysing immersion. In Digital Games Research Association Conference: Changing Views - Worlds in Play

Gow, J., Cairns, P., Colton, S., Miller, P., & Baumgarten, R. (2010). Capturing Player Experience with Post-Game Commentaries. In Proceedings of the 3rd Annual International Conference Computer Games Multimedia Allied Technology. Citeseer.

Gross, J. J., & Thompson, R. A. (2009). Emotion regulation: Conceptual foundations in Handbook of emotion regulation. (J. J. Gross, Ed.) (1st ed.). New York: Guilford Press.

Hellquist, P., Levine, K., & Schyman, G. (2007). BioShock. 2K Games & Feral Interactive (MacOS X).

Ijsseltein, W. A., Poels, K., & De Kort, Y. (2008). The game experience questionnaire: Development of a self-report measure to assess player experiences of digital games: FUGA technical report, Deliverable 3.3.

Jennett, C., Cox, A. L., & Cairns, P. (2008). Being " In the Game ." In Conference Proceedings of the Philosophy of Computer Games (pp. 210–227).

Kikuchi, K. (2012). Fatal Frame II: Crimson Butterfly. Tecmo, Ubisoft, Microsoft Game Studios & Nintendo.

Kleinginna, P., & Kleinginna, A. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. Motivation and Emotion. doi:5:345-379

Lazarus, R., & Lazarus, B. (1994). Passion and Reason: Making Sense of Our Emotions. Oxford University Press.

Lazzaro, N. (2005). Why We Play Games : Four Keys to More Emotion Without Story. Design, 18, 1–8. doi:10.1111/j.1464-410X.2004.04896.x

Minsky, M. (1980). "Telepresence." MIT Press Journals, 45–51.

Moreira, V. H. V. G. (2010). BioStories Geração de Conteúdos Multimédia Dinâmicos Mediante Informação Biométrica da Audiência.

Nacke, L. E., Stellmach, S., & Lindley, C. a. (2010). Electroencephalographic Assessment of Player Experience: A Pilot Study in Affective Ludology. Simulation & Gaming. doi:10.1177/1046878110378140

Nacke, L., & Lindley, C. A. (2008). Boredom, Immersion, Flow - A Pilot Study Investigating Player Experience. In Conference on Game and Entertainment Technologies.

Nakazawa, K., Yamaoka, A., Ito, M., & Owaku, H. (2003). Silent Hill 3. Konami Computer Entertainment Tokyo.

Netter, F. H. (1997). Atlas of Human Anatomy (2nd ed., p. 525). Rittenhouse Book Distributors, Inc.

Noback, C. R., Ruggiero, D. A., Demarest, R. J., & Strominger, N. L. (2005). The Human Nervous System: Structure and Function (6th ed., p. 416). Humana Press.

Ortony, A., Clore, G., & Collins, A. (1990). The Cognitive Structure of Emotions. Cambridge University Press.

Picard, R. W. (1995). Affective Computing MIT Technical Report #321.

Rau, P.-L. P., Peng, S.-Y., & Yang, C.-C. (2006). Time Distortion for Expert and Novice Online Game Players. CyberPsychology & Behavior, 9(4), 396–403.

Ravaja, N., Salminen, M., Holopainen, J., Saari, T., Laarni, J., & Jarvinen, A. (2004). Emotional response patterns and sense of presence during video games: Potential criterion variables for game design. In Proceedings of the third Nordic conference on Human-computer interaction (pp. 339–347). doi:10.1145/1028014.1028068

Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., … Alcañiz, M. (2007). Affective Interactions Using Virtual Reality: The Link between Presence and Emotions. CyberPsychology & Behavior, 10(1), 45–56. doi:10.1089/cpb.2006.9993

Schofield, G., Robbins, B., Ellis, W., Remender, R., Johnston, A., & Graves, J. (2008). Dead Space. Electronic Arts.

Sennett, R. (1993). Authority. W. W. Norton & Company.

Slater, M. (2003). A Note on Presence Terminology. Presence Connect, 3(3).

Slater, M., & Usoh, M. (1994). Body centred interaction in immersive virtual environments. Artificial Life and Virtual Reality, 125–148.

Stearns, C., & Stearns, P. (1989). Anger: The Struggle for Emotional Control in America's History. America's History. The University of Chicago Press.

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). Psychophysiological recording (2nd ed.). New York: Oxford University Press.

Tijs, T. J. W. (2006). Quantifying Immersion in Games by Analyzing Eye Movements. Department of Computer and Systems Science, Royal Institute of Technology, Stockholm (pp. 1–4). Retrieved from http://www.benschweitzer.org/WORK/game_heuristics/Quantifying_immersion_in_games_by_analyzing_eye_movements.pdf

Vorderer, P., & Bryant, J. (2006). Playing Video Games: Motives, Responses, and Consequences (pp. 183–184). Lawrence Erlbaum Associates.

Wood, R. T. A., Griffiths, M. D., & Parke, A. (2007). Experiences of Time Loss among Videogame Players: An Empirical Study. CyberPsychology & Behavior, 10(1), 38–44.

Chapter III

EMOTIONAL RECOGNITION METHODS

HUMAN EMOTION IN INTERACTIVE ENVIRONMENTS: ENSEMBLE AND GROUNDED APPROACHES FOR EMOTIONAL STATE RECOGNITION USING PHYSIOLOGICAL DATA

OUTLINE

Having discussed our conceptual framework for dynamic emotionally inductive experiences and the associated psychophysiological concepts, we now focus on the first step towards emotional regulation; emotional detection. Both identifying the actual effect of emotionally reactive physiological experiences and extracting emotional reaction models from players requires a continuous stream of emotional state data. As such, in this chapter we focus on the implementation of our emotional recognition module (PIERS).

We present a general method for human emotional state recognition in interactive environments. The proposed method employs a three-layered classification process to model the arousal and valence emotional components, based on four selected psychophysiological metrics. Given the real-time requirements of our application, we also present the development of a grounded version, which we will adopt as the de facto implementation throughout the remainder of this thesis.

The modelled emotional states by both approaches compared favourably with a manual approach following the current best practices reported in the literature while also improving on its predictive ability. The obtained results indicate we are able to accurately predict human emotional states, both in offline and online scenarios with varying levels of granularity; thus, providing a transversal method for modelling and reproducing human emotional profiles.

Emotional recognition systems are becoming crucial in many technology areas, ranging from entertainment computing to natural disaster simulations and even phobia or therapeutic treatments. Much research is available on the benefits and necessity of accounting for user's emotional behaviour and responding in an appropriate manner. In more ludic scenarios this may result in an increased immersive experience for the user. It may even be more critical when applied to more "serious" virtual scenarios, such as increasing the effectiveness of therapeutic procedures.

Despite this pressing need, most published work in the current literature relies on various emotional response mechanisms that merely react in a static, hard-coded manner to users' current emotional state.

This is to say that these systems fail to predict users' future reactions and plan their interaction flow accordingly. In turn, this creates two barriers to the dissemination of these techniques: *1)* the creation of fractured body of knowledge that due to most works' narrow focus is difficult to unify and transfer, and *2)* difficulty in evaluating the effectiveness of these implementations in respect to their original objectives.

Given that these systems perhaps lie at the base of a panoply of potentially revolutionary, emotionally-enabled applications (e.g., emotionally-adaptive movies, the evaluation of a therapeutic treatment's efficacy over time or, eventually, virtual emotionally-sentient AIs (Cavazza, Pizzi, Charles, Vogt, & André, 2009; Leite et al., 2010)), we believe there is a need to develop methods that allow us to: *a)* recognise human emotions as accurately as possible so we can theoretically *b)* model human emotional reaction profiles by monitoring human emotional states. In essence, this is one of the cornerstone objectives of this thesis so the development (and validation) of such a method lies at a critical juncture; one on which the success of our research work is dependant.

A reliable, real-time emotional detection system with low calibration overheads

To this end, we are interested in building a general, accurate approach at detecting human emotional states that can easily be incorporated into our conceptual architecture. It should also require a calibration overhead as minimal as possible to avoid prolonging experimental sessions for too long, thus contaminating participants; and allow an easy integration into existing workflows.

Following this objective, this chapter is structured as follows: Section 3:1 describes the current state of the art in emotional recognition. As they are fundamentally different, it divides this discussion in three categories: subjective, objective and gameplay-based modelling. It concludes by presenting a series of applicability considerations. Following this discussion we present a concise but thorough primer on the employed physiological metrics in section 3:2. In this section we also substantiate our choice of physiological metrics and modelling approach. From this point onwards we present our approach at physiological modelling of human emotion from an interactive environment perspective. In section 3:3 we discuss the chosen experimental conditions and procedure, as well as feature extraction. Throughout section 3:4 we present our main contribution; the adopted methodology in our approach. These range from unifying the collected metrics across various participants, to the creation of data fusion classifiers to combine the metrics, to leveraging the various classifiers' predictive abilities in two ensemble approaches. We conclude section 4:4 by also presenting a grounded version of our main approach for use in

real-time and low (experimental) overhead contexts. In section 3:5 we validate both the aforementioned approaches by comparing their performance to a manual approach using the current literature's accepted physiological modelling practices. Section 3:6 discusses the obtained results, commenting on their performance and possible future improvements. Finally, section 3:7 presents a brief summary of the contributions contained throughout this chapter and contextualises the logical progression towards the following chapter.

3.1 IDENTIFYING AND RECOGNISING HUMAN EMOTIONAL STATES

A comparison of emotional state detection taxonomies

Various taxonomies exist for emotional detection systems in a wide range of application domains, each one with its own dedicated literature. Within our current context – video games and digital media – Yanakakis et al. segment these taxonomies into three distinctive types of approach: Subjective, Objective and Gameplay-based (Yannakakis & Togelius, 2011). Although the authors refer to them as types of player experience modelling, the base concept is that these approaches attempt to define how each player affectively interprets the gaming experience and is thus mainly a matter of nomenclature.

Subjective Modelling

Subjective player experience modelling (SPEM) resorts to first-person reports on the player's experience. While the most descriptive of the three, it is difficult to properly analyse since reports tend to be plagued by experimental noise derived from player self-deception effects, memory limitations and intrusiveness; for example if questionnaires are performed during the experience itself (Yannakakis & Togelius, 2011). However, when properly timed and framed, data collected through these methods can provide powerful ground truth data for more data-driven techniques.

Objective Modelling

Objective player experience modelling techniques (OPEM) attempt to explore the possible correlations between game events and physiological alterations. This approach usually employs multiple input modalities for real-time applications (Yannakakis & Togelius, 2011). Objective modelling techniques are further divided into two types: model-based and model-free. While model-based approaches link physiological changes to popular models derived from emotion theories, such as for example, Russell's continuous arousal and valence (pleasure)

dimensions (Russel, 1980a) or Plutchik's Emotion Wheel (Plutchik, 1980) (see Figure 3:1), model-free techniques build their mappings based solely on user annotated data - usually to discrete emotions (e.g., fear, surprise, anger). However, systems may not rely solely on either of these two types. Hybrid approaches assume some type of correlation exists between physiological measures and affect (i.e., assume a pre-existing model), but seek to define the structure of the correlation function by analysing the available data. In fact, many known systems use the latter approach; they assume a theoretical model of emotion as their structure and build the mappings via the annotated data (Chanel, Kronegg, Grandjean, & Pun, 2006; Drachen, Nacke, Yannakakis, & Pedersen, 2010; Mandryk & Atkins, 2007; Moreira, 2010).

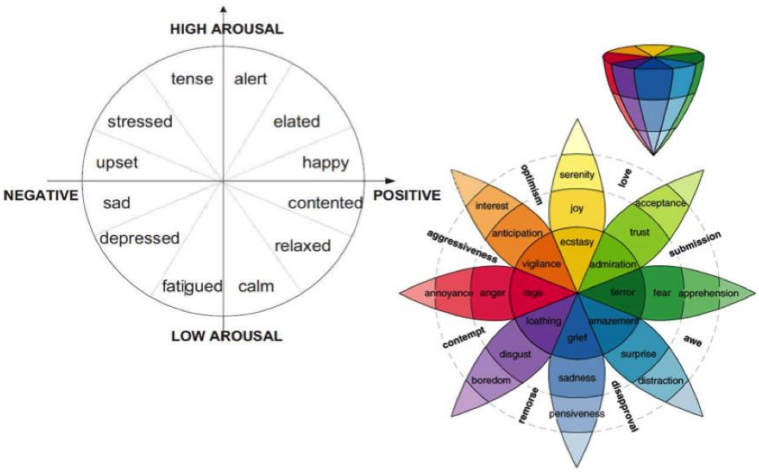


Figure 3:1. Russell's arousal-valence space (left) and Plutchik's emotion wheel (right). Sections on the AV space show how the two dimensions can be used to identify more complex emotional constructs such as, for example, fear or sadness. (Adapted from (Russel, 1980) and puix.org, respectively).

Various successful attempts have been made in the field of emotion recognition using the aforementioned types of objective modelling techniques – the hybrid type being clearly the most popular one. For instance, (Chanel, Kronegg, Grandjean, & Pun, 2006) was able to classify arousal using naïve Bayes classifiers and Fisher Discriminant Analysis (FDA), based on Electroencephalogram (EEG), skin conductance (SC), blood volume pressure (BVP), heart rate (HR), skin temperature (ST) and respiration rate measures. Complementary to this work, Leon et al. propose the classification of valence in three intensity levels, using similar SC and HR measures and auto-associative neural networks (Leon, Clarke, Callaghan, & Sepulveda, 2007). Similarly, Brown et al. propose a K-Nearest Neighbour approach at detecting valence using frontal alpha wave asymmetry indices in

EEG readings (Brown, Grundlehner, & Penders, 2011). In a more direct approach, Hazlett has found facial EMG can also be used to distinguish valence in gameplay experiences (RL Hazlett, 2006), having then further expanded these findings towards general software experiences (R Hazlett & Benedek, 2007). More recently, Nacke has proposed practical recommendations on the usage of physiological player metrics from evaluating games, thus further establishing their use in everyday scenarios (L. E. Nacke, 2013).

In terms of simultaneous arousal and valence detection, (Haag, Goronzy, Schaich, & Williams, 2004) propose employing EMG, SC, ST, BVP, ECG and respiration rates to classify emotional states, reporting 89% accuracy for arousal and 63% for valence, with a 10% overall error margin. Under similar circumstances, Nasoz et al. have also successfully classified more complex emotional constructs, such as "happiness" and "fear" using a multi-modal approach (Nasoz, Lisetti, Alvarez, & Finkelstein, 2003).

Real-time, continuous approaches at emotional state modelling

Within the proposed line of low calibration approaches, the work by (Vinhas, Silva, Oliveira, & Reis, 2009) proposes a system capable of measuring both arousal and valence in real-time, using the subject's SC response and HR derivative. A key factor of this work is that it introduced a continuous classification of arousal and valence. However, the method not only uses a limited set of psychophysiological measures, which limit its coping abilities to unseen scenarios, it also does not present sufficiently convincing results for valence classification using the HR derivative. Furthermore, as with the remaining literature, the issues arising from inter-subject physiological variations (see section 3:3) are not effectively solved (although they are acknowledged).

Finally and similar to Vinhas, Mandryk presents an approach based on fuzzy logic that classifies EKG, EMG and SC measurements in terms of both arousal and valence (R. Mandryk & Atkins, 2007; Vinhas et al., 2009). Due to the ambiguous nature of physiological data, fuzzy logic presents itself as an elegant solution.

Gameplay-Based Modelling

According to Yanakakis et al., gameplay-based player experience modelling (GPEM) is driven by the assumption that the actions of users partaking in an interactive environment and/or their preferences regarding these actions or environment are also linked to his affective experience. This leads to the belief that such preferences can be used in detriment of more intrusive measures to identify the user's affective experience (Yannakakis & Togelius, 2011). In essence, the rationale is virtually the same as in objective modelling. The annotated data

obtained from the player's gameplay session is analysed and interpreted in light of some cognitive or behavioural model or theory. Since these approaches are not on our research's focus, we will limit their discussion. It is also important to note that while this type of modelling is the least intrusive of all three, it has been noted to result in low-resolution models of players' affective experience and models are often based on several strong assumptions between player behaviour and preferences (Yannakakis & Togelius, 2011).

In the field of gameplay-based modelling techniques, (Leite et al., 2010) has presented an empathic robot capable of reacting to the user's affective state. This affective state is inferred through contextual information collected from the surrounding environment and interpreted according to an empathic behaviour model. Using the same approach, complex game aspects such as a game's narrative have also been shown to be dynamically adaptable to individual players, in such a way that a pre-determined reaction is achieved (Figueiredo & Paiva, 2010). Along the same research avenue, (Pedersen, Togelius, & Yannakakis, 2009) have also shown the feasibility of constructing offline computational intelligence models capable of predicting optimal game parameter sets for the elicitation of certain affective states.

Applicability Considerations

Although varied in their application focus, these approaches act as proofs-of-concept for the feasibility of real-time emotion detection systems in affective computing applications. However, when taking into consideration the previous discussion on emotional modelling, it becomes apparent that building a generic or quasi-generic emotional state detection system that also requires minimal calibration is considerably harder for subjective and gameplay-based modelling techniques.

On one hand, SPEM techniques' noisy data collection and high subject and scenario dependency make it a laborious approach. On the other hand, GPEM techniques are easier to generalise due to the re-utilization of extracted features. However, they in turn require a high amount of domain-specific knowledge, which is almost always coupled with strong assumptions and fairly complex adaptations in each new implementation. Furthermore, preferences must be re-learned for each new scenario, given that they may not always be consistent.

Hybrid OPEM techniques provide the best of both worlds

OPEM techniques are the most popular approach because data inputs and signal processing methods are fairly consistent between applications and - given the independent theoretical models - data

interpretation is also maintained (i.e., minimal domain knowledge is required). As an added bonus, Russell's AV space offers a continuous model of emotional states on which it is easy to represent emotional transitions and superimpose more complex emotional constructs such as fear or joy. Given these characteristics and its computationally-friendly nature, it has been the most widely adopted model.

In light of these conclusions, we have chosen a hybrid OPEM approach based on skin conductance, corrugator supercilii (brow) and zygomaticus major (cheek) facial electromyography and electrocardiographic metrics, represented in terms of arousal and valence on Russell's AV space.

3.2 SELECTION OF PHYSIOLOGICAL METRICS

Based on the discussed literature, we conducted a survey on the most successful physiological channels and features. In decreasing order of importance, four main factors were taken into account:

Emotional recognition system requirements

(1) *Precision:* How accurately can the signal be interpreted in terms of either arousal or valence?
(2) *Sensor Reliability:* How likely is the sensor to fail and how much calibration does it require?
(3) *Signal Processing*: How much signal processing (e.g., filtering, noise reduction) does the channel require in order to allow extracting meaningful features, and is this processing cost affordable in a real-time scenario?
(4) *Intrusiveness*: Would the required apparatus interfere with the gameplay experience, potentially biasing it in any significant way?

This survey led us to adopt skin conductance, facial electromyography and electrocardiography-based metrics. All of these channels have become popular in the affective computing literature by providing the most accurate and interpretable measurements. A brief description of each metric follows.

Electrodermal Activity

Electrodermal activity (EDA), usually measured in the form of skin conductance (SC), is a common measure of skin conductivity. EDA arises as a direct consequence of the activity of eccrine (sweat) glands. Some of these glands situated at specific locations (e.g., palms of the hands and feet soles) respond to psychological changes and thus EDA/SC measured at these sites reflects emotional changes as well as

cognitive activity (Stern et al., 2001). SC has been linearly correlated with arousal (Chanel et al., 2006; R. Mandryk & Atkins, 2007) and extensively used as stress indicator (Vinhas et al., 2009), in emotion recognition (Chanel et al., 2006; RL Hazlett, 2006; Leon, Clarke, Callaghan, & Sepulveda, 2007; R. Mandryk & Atkins, 2007; Vinhas et al., 2009) and to explore correlations between gameplay dimensions (Pedersen et al., 2009). In our experiment, we measured SC using two Ag/AgCL surface sensors snapped to two Velcro straps placed around the middle and index fingers of the non-dominant hand (Stern et al., 2001).

Cardiovascular Measures

The cardiovascular system is composed by the set of organs that regulate the body's blood flow. Various metrics for its activity currently exist, among which some of the most popular ones are: blood pressure (BP), blood volume pulse (BVP) and heart rate (HR). Deriving from the HR various secondary measures can be extracted, such as inter-beat interval (IBI) and heart rate variability (HRV). HR is usually correlated with arousal (Stern et al., 2001). HR, along with its derivate and HRV has also been suggested to distinguish between positive and negative emotional states (valence) (R. Mandryk & Atkins, 2007; Vinhas et al., 2009). In our experiments HR and HRV were inferred from the participants' BVP readings using a finger sensor.

Electromyography

Electromyography (EMG) is a method for measuring the electrical potentials generated by contraction of muscles (Stern et al., 2001). Facial EMG has been successfully used to distinguish valence in gameplay experiences (RL Hazlett, 2006). In the former experiences, Hazlett describes the zygomaticus major (cheek) muscle as significantly more active during positive valence events and the corrugator supercilii (brow) muscle as more active in negative valence events. We measured facial EMG through surface electrodes on the corrugator supercilii and zygomaticus major muscles (see Figure 3:2).

3.3 EXPERIMENTAL DETAILS

To gather enough ground truth data to determine whether we could build an emotion recognition system for gameplay and general multimedia content, we conducted a series of controlled experiments with a total of twenty-two healthy participants. In line with the practices presented in the related literature (Chanel et al., 2006; Haag

et al., 2004; Leon et al., 2007; R. Mandryk & Atkins, 2007; Stern et al., 2001; Vinhas et al., 2009), the applied protocol was initially tested and refined in an iterative prototypical cycle, using several pilot studies comprising a total of 10 participants. The results reported in this chapter apply to the data collected and processed for the remaining twelve participants in the final iteration of the experimental procedure. Participants ranged from undergraduate students to more senior researchers and were aged 22 to 31 (M=24.83, SD=2.29), which naturally constitute the demographic we limit our findings to.

Experimental Conditions

Each session was divided into three conditions designed for obtaining the necessary data samples to train our system. The first two conditions were aimed at eliciting extreme arousal and valence ratings: the first one being a 10-minute long session of relaxing music and the second one playing the horror video game Slenderman, by Parsec Productions. The third condition was aimed at eliciting neutral to mild reactions using a set of 36 images from the IAPS library (Lang, Bradley, & Cuthbert, 2008), representative of its full gamut (i.e., low to high elicitation potential) - refer to Figure 3:2 for an illustration of the used material.

Participants, conditions and collected data

In each of the experimental conditions, the participant's SC, facial EMG and BVP readings were recorded (Figure 3:2). SC was measured at the subject's index and middle fingers using two Ag/AgCL surface sensors snapped to two Velcro straps. BVP was measured at the thumb using a clip-on sensor. Both these readings were made at the non-dominant hand (Stern et al., 2001). Facial EMG was measured at the zygomaticus major (cheek) and the corrugator supercilii (brow) muscles and correlated with positive and negative valence, respectively (Stern et al., 2001). Sessions were timed and conducted in a room with acoustic insulation and controlled lighting and temperature conditions. Participants were isolated in the room at the beginning of each section in order to limit contamination effects. The only human interaction was during the relaxation/briefing periods in between conditions.

Experimental Protocol

All participants were exposed to each condition in a fixed order. During our pilot studies, we found it was necessary expose the participants to the music and the horror videogame conditions before presenting the IAPS images; otherwise they tended to rate the images relatively to one another, instead of on an absolute scale. By using the relaxing music and video game to delimit their responses, participants were able to

rate the IAPS images in more absolute terms and a drastic reduction (~40%) in the ratings' standard deviation was indeed observed.

After signing an informed consent form, participants were briefed and underwent the full experimental protocol. Each condition was preceded by a relaxation period of approximately 5 minutes, through which baseline (averaged) values for each channel were extracted. The participants then underwent each of the experimental conditions, whilst reporting their affect ratings.

A wide range of content improves contextualisation of participants' emotional reactions

Figure 3:2. Experimental conditions and procedure. Top-left: Sensor placement on a participant using headphones for noise cancellation. Top-right: Screenshot of the physiological data collection. Bottom-left: Screenshot of an encounter with the Slenderman monster. Bottom-right: Illustrative sample from the IAPS library.

Experimental conditions were sub-divided into its constituent training samples, each with the same length as the event, plus a buffer length of 5 seconds added to both its boundaries. Regarding training samples, since we were interested in the participant's lowest emotional activation values for the relaxing music condition, in this case the training sample was equal to the whole condition. On the remaining two conditions, each event – images for the IAPS condition and gameplay events for the Slenderman condition – was isolated and independently rated by the participants.

Regarding the annotation procedure, participants rated the training samples immediately after their presentation. The exception to this was

the Slenderman condition, since interrupting the gameplay activity to rate each event was not only intrusive; it also implied our physical presence, which could contaminate the experience. As such, for this condition, the gameplay session was recorded by a commercial frame grabber (Fraps by Beepa Pty Ltd.) and analysed in conjunction with the participant in a post-gameplay interview. Participants reported their absolute maximum arousal and valence ratings in a 21-point Likert scale questionnaire ranging from -5 to 5 in 0.5 increments. We chose to ask participants to rate each event according to their absolute maximum because it introduced the least noise in the annotation process since it is harder for individuals to rate their average affect over a 10 or 20-second time window than to isolate a more intense emotional peak.

Each condition had an average duration of 10 to 12 minutes, with the exception of the terror videogame, which usually took 15 to 20 minutes. Overall, from setup to debriefing, the experiment had an approximate duration of 2 hours.

Apparatus

All of the sessions were performed on a laptop computer running Windows 7. The monitor was a 17" LCD display running at a resolution of 1920x1200 pixels. The gaming condition was recorded and synched with the physiological data at 30 Hz, using their starting timestamps (see Figure 3:2). Physiological data was collected and exported using the Nexus-10 hardware by Mind Media.

Data Analysis & Feature Extraction

Regarding data analysis and feature extraction, the raw HR, heart rate variability (HRV) and SC readings were collected at 32Hz, while facial EMG was collected at 2048 Hz. HR and HRV (R-R intervals) readings were computed from the raw BVP readings. All of the physiological signals were then exported to a tab-delimited text file sampled at 32 Hz using the BioTrace+ software suite for future analysis.

Due to past hardware failures, the exported raw data was filtered for anomalies. Since no corrupt data was observed all of the collected data was retained for analysis. Regarding data pre-processing, HR readings were filtered using an irregularity detection method similar to (Haag et al., 2004). This method only allowed a 25% variation of the HR signal at each new reading, based on its past variation over the previous 5 seconds. HRV was then recomputed from the corrected HR values. EMG amplitude values were extracted from the raw EMG readings

using the Root-Mean Square procedure. Raw readings were corrected by subtracting their corresponding baseline values. Subsequent signal analysis revealed no additional filtering was necessary. Finally, sensor readings were smoothed over a moving window. Initially we employed an approximated Gaussian kernel for this smoothing process. However, this introduced unwanted signal distortions and was thus replaced with a traditional moving average kernel. Window sizes for HR and HRV were 2 seconds, 5 seconds for SC and 0.125 seconds for EMG, as suggested in (Stern et al., 2001).

As previously mentioned, each condition was segmented into several training samples that were independently rated by participants using a 21-point Likert scale that ranged from -5 to 5. These were later converted to the 0-10 range (the same range as the IAPS ratings). Since sometimes participants had difficulty in differentiating similar stimulus using only the Likert scale, they were also allowed to provide a numeric answer, using the Likert scale as a reference. This implies that the gathered ground truth data can be treated as numeric, rather than nominal. Also, since participants were asked to rate their absolute emotional peaks, so were these values extracted from each channel's training sample.

The amount of collected data largely exceeds the literature's mean

Overall, an average of 230 data points were collected per participant: the minimum values for the relaxing music condition (5 data points, one per channel), 36 samples for the IAPS images condition (180 data points), an average of 6 samples (30 data points) in the terror videogame condition – number of gameplay events varied across participants – and 3 neutral baseline samples (15 data points).

3.4 DETECTING AV STATES

This section details how the annotated ground truth gathered in the three experimental conditions was used to detect the participants' emotional states. The developed method detects participants' arousal and valence ratings through a three-layer classification process. The first classification layer simultaneously scales and correlates each input to an AV (arousal-valence) dimension by applying participant-specific regression models (i.e., each individual has his own set of regression models – one per physiological metric/AV correlation). This regression process generates a numeric output that is then fed to the second classification layer, which combines these predictions through various fusion classifiers into one final rating by minimising their intrinsic error margins. Since each fusion classifier (the term is used with a loose connotation throughout this chapter) uses different classification

The classification process summarised

rationales to fuse the regressed metrics, the third layer leverages this knowledge through an ensemble model that weighs each classifier's prediction based on their mean absolute error (MAE) value. The full process is depicted in Figure 3:3. Each of the described layers will be discussed in detail throughout this section.

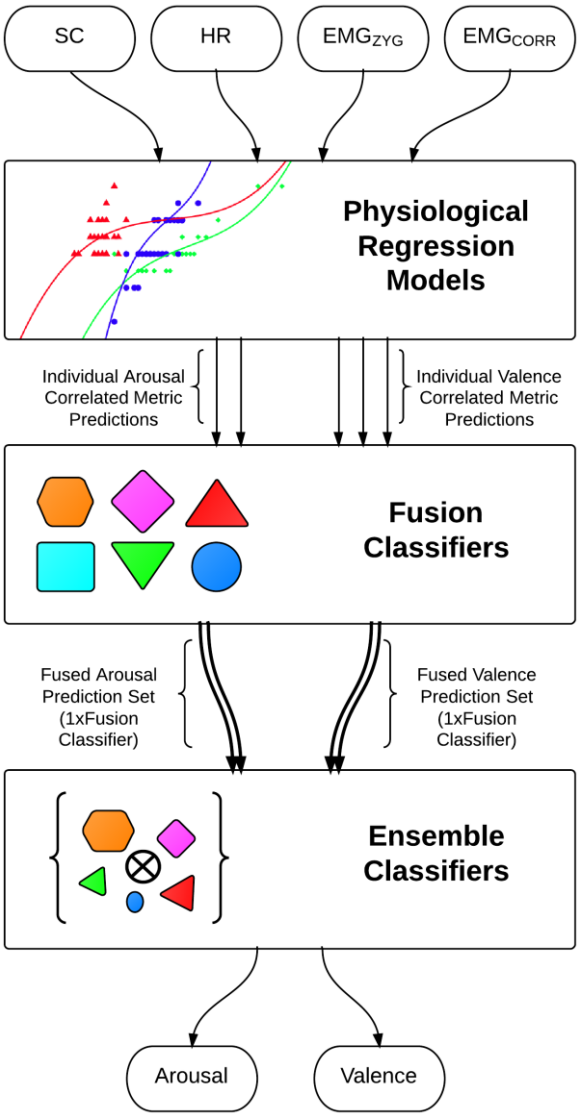


Figure 3:3. Proposed architecture. Each of the four physiological metrics is used to create two distinct sets of arousal-correlated predictions. These predictions are then fed in parallel to several machine learning (fusion) classifiers, which combine them using different classification rationales. These classifications are ultimately weighted according to the fusion classifiers' MAE values into a final arousal prediction. An identical process is used to estimate valence.

Physiological Input Regression Models

One of the most common issues with emotional recognition systems is the difficulty in obtaining an absolute scaling for the reported measures, an open issue that has been well documented in the literature (Levillain, Orero, Rifqi, & Bouchon-Meunier, 2010; R. Mandryk & Atkins, 2007; Vinhas et al., 2009). This usually occurs due to either: *a)* the used material's inability to elicit meaningful emotional alterations, or *b)* by failing to annotate how each recorded event actually impacted the participant (i.e., assuming that a full emotional range elicitation occurred throughout the process). This issue is further aggravated by the considerable differences in physiological ranges, baseline values and activation functions displayed across individuals.

Regressing the obtained data provides a more trustworthy model of players' AV ratings

We tackled the insufficient emotional elicitation aspect of the scaling problem by exposing participants to a wide gamut of emotional content; ranging from a relaxing music session, to a sizeable, representative set of visual stimuli from the International Affective Picture System (IAPS) library (Lang et al., 2008), and the psychological terror videogame, Slenderman. The second aspect of the scaling issue (unduly annotation) was addressed with a carefully planned experimental protocol – as discussed in detail throughout section 3:3.

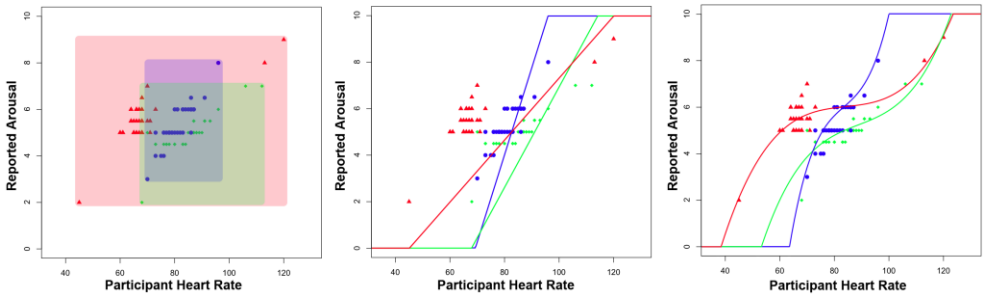


Figure 3:4. Effect comparison between normalising and regressing physiological metrics (in this case, heart rate). Left: Annotated data points for three participants (red, green and blue regions/dots). Middle: Correlation functions produced by blindly normalising the obtained data for each participant. Right: Correlation functions produced by fitting a polynomial (3rd degree) regression model to the same annotated data.

The most common method of addressing both aspects of this issue – insufficient emotional elicitation and inter-participant physiological variations – is by normalising the recorded values after subtracting a baseline value. However, this not only assumes that all participants experienced the same emotional ranges but, more importantly, that all of them experienced these ranges to their fullest extent. Thus, this practice introduces a fairly high amount of noise into the classification process by forcing the assumption of improbable correlation functions

(see Figure 3:4). We approached this problem from a different perspective. Instead of a simple baseline subtracted-normalisation, we explored the correlation functions between each of the physiological metrics and the subjective ratings. For this effect and taking into consideration that the used physiological metrics are described in the literature as fairly robust indicators of arousal and valence, respectively, we chose to use regression models. Regression analysis has also been successfully used in emotional reaction identification to musical (Yang, 2008) and audio-visual stimuli in PET scans (S et al., 2005), thus further motivating our choice. A visual comparison between our approach and the traditional method is presented in Figure 3:4. Notice how the normalisation method produces much higher errors, especially in the extreme regions of the spectrum. The regression method, however, minimises these errors through its least-squares fit. It is also easy to overfit the data using the regression method so countering measures should be taken (e.g., using bi-directional stepwise regression to determine the least complex, significant model).

By using the participants' own subjective ratings as our ground truth we were able to simultaneously take into account each participant's individual physiological activation functions and link them to the AV dimensions. We proceeded to explore the correlations of each annotated physiological channel to the AV dimensions and – apart from the HRV channel – confirmed each of the correlations referred in the literature. However, despite HR negatively correlating with valence, this was not observed for all three experimental conditions, since HR did not significantly fluctuate in the IAPS condition. As such, the results reported in Table 3:1 for the HR-valence correlation refer only to the training samples extracted in the first two experimental conditions, not all three as per the remaining described correlations. An interesting aspect to this correlation is that while HR has been positively correlated to valence in previous work (Drachen, Nacke, Yannakakis, & Pedersen, 2010; R. Mandryk & Atkins, 2007), our results showed an inverse correlation between HR and valence. In fact, this was to be expected because our scenario (a psychological horror game) is fundamentally different from the ones in previous works (sports (Regan Lee Mandryk, 2005) and first-person-shooter games (Drachen et al., 2010)). In a horror game a higher HR usually meant the player triggered a scare event and thus generated a low valence response, whereas in a sports or first-person-shooter game events tend to have more positive connotations (e.g., scoring a goal or killing an enemy). This seems to suggest a certain degree of interplay between arousal and valence, indicating that while HR may be a reliable fallback to predict valence, the correlation must be verified for each new scenario. This phenomenon is also hypothesised in

Confirming and adapting previously observed correlations between emotional states and human physiology

(Drachen et al., 2010; Regan Lee Mandryk, 2005); an hypothesis to which our findings seem to give strength.

In this exploratory phase we used both linear and non-linear (third degree polynomial) models. Model complexity was kept in check using bidirectional stepwise regression. The procedure was based on their adjusted-R^2 values in order to minimise the effects of a large number of predictors on the polynomial models. The selected models were evaluated for correctness using cross-validation and can be inspected in Table 3:1.

Table 3:1. Fitness values for the created regression models. Model complexity was controlled using stepwise-regression.

Metric	Dimension	Adjusted-R^2 Model Values (μ, σ)	
		Linear	*Polynomial*
SC	Arousal	$(0.90 , 3.8^{-2})$	$(0.95 , 3.0^{-2})$
HR	Arousal	$(0.68 , 7.1^{-2})$	$(0.74 , 8.9^{-2})$
EMG_{Zyg}	Valence	$(0.84 , 1.4^{-2})$	$(0.92 , 1.6^{-1})$
EMG_{Corr}	Valence	$(0.83 , 7.9^{-2})$	$(0.95 , 7.5^{-2})$
HR	Valence	$(0.88 , 1.0^{-1})$	$(0.96 , 6.4^{-2})$

Statistically determining minimal necessary regression model complexity

Although there are multiple accounts of a linear correlation between SC and arousal (see section 3:1), there is no common evidence that the remaining collected metrics correlate linearly with any of the AV dimensions. Thus, a statistical analysis between the generated linear and polynomial models was conducted and revealed that non-linear correlations are indeed supported by our data. One-tailed paired t-tests using the models' adjusted-R^2 values as within-subject conditions revealed statistically significant ($p < 0.05$) differences between the linear and polynomial models for the following correlations: SC-arousal (t = -2.397, p = 0.035), HR-arousal (t = -2.393, p = 0.036), EMGZyg-valence (t = -2.396, p = 0.038), EMGCorr-valence (t = -2.825, p = 0.018) and HR-valence (t = -3.297, p = 0.007). Closer inspection also revealed that although the polynomial SC-arousal models presented a significant improvement over the linear ones, they were marginally different from the latter and only presented a small fitness improvement (5%), while the remaining models presented improved fitness values ranging from 9 to 14%. We thus decided to maintain the linear model for SC-arousal and opt for the polynomial models in the remaining ones.

AV Fusion Models

Since the previous classification stage generates an overlapping set of ratings for each AV dimension, it became necessary to fuse these sets in order to obtain a unified AV prediction. Given that we were aiming at a hybrid approach between a theoretically-grounded and ML-based method and had already established the physiological mappings to our theoretical model's dimensions, we chose to leverage the benefits of a more data-driven technique: the ability to find more complex relationships in the data and better generalization capabilities across multiple subjects without prior strong assumptions on the data's structure or properties (Yannakakis & Togelius, 2011).

Using fusion models to leverage the strengths of each physiologic channel

The hypothesis behind this step was that the fusion models would be able to leverage each channel's error functions to decide how their predictions should be combined in the final rating. In other words, it would be able to optimize the combination policy for the regression models' ratings, according to their individual characteristics.

Multiple ML classifiers were trained using the regressed data. Classifier diversity was promoted in the selection process with the intention of exploring various classification logics - and further benefiting from them (see following sub-section). The selected classifiers were: decision trees, single-layer perceptron neural networks, random forests and support vector machines.

Regarding decision trees, the splitting criterion was determined using the normalised information gain at each node and trees were pruned to avoid over-fitting using the expected reduction in the Laplace error estimate. Neural networks were parameterised with a hidden layer comprised of 10 neurons and trained using a back-propagation algorithm with a stopping threshold of 0.1 units. It has been suggested that while random forests are a "random" statistical method and thus do not require cross-validation techniques to obtain unbiased results, it is advisable to test their stability by repeating the training phase with an increasing number of trees until convergence is met (Strobl, Malley, & Tutz, 2009). Thus, we trained models for each AV dimension with a number of trees ranging from 50 to 5000, following a regularly sampled, linearly increasing function and chose the models where convergence was attained. This was at 500 trees for arousal and 2000 trees for valence. The number of randomly preselected predictor variables used at each split was the square root of the number of predictor variables, also as suggested in (Strobl et al., 2009). Finally, the support vector machine classifiers were trained using three kernel types from the 'e1071' R library: a linear, a polynomial and a radial kernel. Gamma

and epsilon values were maintained at their default values (0.3 and 0.1, respectively). All classifiers were taught to rate the provided inputs (the regressed physiological data) according to the ground truth provided by the participants.

A final contemplation regarding these models is that, as previously mentioned, since the regression models have already accounted for the participants' own physiological traits, the models employed in the second classification layer are not participant-dependent. Thus, the amount of training data available for these methods was substantially higher. To test the accuracy of the built models, they were validated using 3-fold and 10-fold cross validation techniques. Since folds were already pre-calculated individually for each participant in the first layer, the folds for each classifier were computed by randomly merging one fold from each participant to generate the "population" folds. All of the presented classifiers were trained and validated using these same folds. While this served no significant computational purpose or gain, it avoided injecting unseen data into the second layer's training samples. Care was also taken to, as much as possible, equally divide the training samples across classes, so as to not bias the classifiers.

Evaluating performance and stability of each created model type

Table 3:2.1. Regressed arousal fusion classifier accuracy ratings (%). The final row also presents each classifier's mean absolute error (MAE) as an overall indicator of its performance.

Error Threshold	Arousal Fusion Classifiers (3-fold, 10-fold CV)						
	DT		NN		SVM		RF
0.2	57.0	57.0	64.4	67.4	56.5	63.0	71.5
0.5	84.5	89.0	97.4	95.5	95.7	95.7	91.3
1.0	98.2	99.0	99.1	100	99.1	99.1	98.0
MAE	0.26	0.26	0.38	0.17	0.22	0.19	0.18

Table 3:2.2. Regressed valence fusion classifier accuracy ratings (%). The final row also presents each classifier's mean absolute error (MAE) as an overall indicator of its performance.

Error Threshold	Valence Fusion Classifiers (3-fold, 10-fold CV)						
	DT		NN		SVM		RF
0.2	69.0	64.3	63.5	52.4	56.7	50.0	72.5
0.5	84.6	76.2	91.3	71.4	88.5	76.2	84.8
1.0	92.3	85.7	95.2	85.7	91.3	88.1	92.0
MAE	0.35	0.61	0.34	0.46	0.35	0.41	0.26

The obtained results (Tables 3:2.1 and 3:2.2) show that we were able to identify arousal with as much as 97% accuracy and valence as precisely as 91%, using neural networks, with an acceptable error margin of 0.5 points in the AV scale. Notice that given the data's

distribution (see Figure 3:4), a 1.0 error should not be considered as acceptable as it overly simplifies the classification problem. This means it serves merely to contextualize the obtained results in regards to the grounded and manual approaches (see sections 3:4 and 3:5). As such, in this phase only the 0.2 and 0.5 error thresholds are of significant importance. The presented results indicate that the fusion classifiers were able to accurately classify arousal and valence with 0.2 and 0.5 error margins, respectively. This indicates that, as expected, valence is harder to classify. Further considerations on the presented results can be found in the last paragraphs of the following sub-section. Each predicted AV rating was evaluated through the binary thresholding function expressed in (Eq. 4:1):

$$T(p) = \begin{cases} 1, & if\ |p - \tilde{g}| \leq t \\ 0, & if\ |p - \tilde{g}| > t \end{cases} \qquad \text{(Eq. 4:1)}$$

Where p is the predicted AV rating by the classification function, $\tilde{g}$ is the ground truth data for the data sample under evaluation and t is the maximum acceptable error threshold for p to be considered correct. Following the observed error margins in the literature, t was set at values between 0.1 and 1.0. The average classification error (final line of Tables 3:2:1 and 4:2:2) was computed as the average absolute difference between each of the predicted values P = p_1, p_2, .., p_n and their corresponding ground truth annotations $\tilde{G}$= $\tilde{g}_1, \tilde{g}_2$, .., $\tilde{g}_n$. We consider that this range of threshold values provide a good overview of how the method performs with varying levels of granularity and, as such, represent its adequacy for the considered scenarios.

Fusion Classifier Ensembles

Having analysed the individual performances of the employed classifiers, we noticed that promoting classifier diversity had a secondary effect besides the observed accuracy ratings - the classification functions were clearly quite different in some cases.

Using ensembles to promote a "deep" learning taxonomy, thus benefiting from the fusion classifiers' more abstract feature sets

Thus, we were curious whether it was possible to leverage this feature to further improve our classification accuracy and turned our attention towards ensemble models. Ensemble models are often referred to in the literature as a method to combine multiple hypotheses to create a better hybridised hypothesis and while the term ensemble usually refers to a method where the same base learner is used, it has spread to encompass the usage of multiple classifiers (Rokach, 2005).

Since the fusion classifiers were not levelled in terms of performance, we decided to adopt two ensemble voting schemes. The

first one was a typical averaging scheme of the classifier's mean absolute error (Eq. 4:2), to be used as a baseline:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - \tilde{g}_i| \qquad \text{(Eq. 4:2)}$$

The second one was an inverted weighted scheme (Eq. 4:3 and Eq. 4:4) also based on the classifier's mean absolute error:

$$\psi_i = \frac{MAE_i}{\sum_{j=1}^{n} MAE_j} \qquad \text{(Eq. 4:3)}$$

$$w_i = \begin{cases} 1 & if\ MAE_i = 0, \\ \dfrac{\left(\frac{\sum_{i=1}^{n} MAE_i}{\psi_i}\right)}{\sum_{j=1}^{n}\psi_j} & otherwise. \end{cases} \qquad \text{(Eq. 4:4)}$$

Determining how to weigh each classifier's contribution is key in achieving a well-balanced, stable prediction

Where ψ_i is the i^{th}'s classifier contribution towards the cumulative mean absolute error of the ensemble's classifier set and w_i is the weight of the same i^{th} classifier, which is attributed through an inverse score normalised through all classifiers, based on their individual ψ value. In other words, w is a proportional inverse measure of the classifiers' misclassification rate - this means the higher the error rate, the lower the classifier's weight towards the ensemble prediction.

Each of these voting schemes was applied to two classifier sets; one containing all of the previously trained classifiers, and one containing only the highest ranking version of each classifier type. The obtained results for each ensemble type and classifier set for both arousal and valence are visible in Table 3:3. The presented results show that the ensemble approach produced improved results. As could be expected, given their sometimes-significant performance variation, weighting the classifiers' votes produced the best results. Likewise, favouring the inclusion of the subset of best classifiers in the ensemble also improved the accuracy ratings, with the optimal-subset weighted ensemble (OWE) achieving the best results on both arousal and valence classification.

Comparing performance improvements between weighting schemes

In sum, using the OWE approach we were able to reduce the mean absolute error on arousal classification from 0.28 to 0.13, while maintaining a classification accuracy of 97.4%. We were also able to improve valence classification by reducing the mean absolute error from 0.4 to 0.32 and, at the same time, actually benefiting from a minimal raise from 91% to 92.3% accuracy. Given that neural networks obtained the best results, since these are highly volatile (as indicated by their

higher MAE) and more prone to overfitting than other methods, we feel that our ensemble approach poses a significant improvement in terms of stability and reliability.

Table 3:3. Accuracy ratings for the arousal ensemble classifiers (%). Best ensemble version highlighted in grey. The final row also presents each classifier's mean absolute error (MAE) as an overall indicator of its performance.

Error Threshold	Arousal Ensemble Classifiers (3-fold, 10-fold CV)							
	Optimal Classifier Set				Full Classifier Set			
	Averaged Voting		Weighted Voting		Averaged Voting		Weighted Voting	
0.2	70.4	73.9	77.4	82.6	54.8	65.2	73.9	80.4
0.5	95.7	93.5	97.4	95.6	94.8	93.5	96.5	95.7
1.0	99.1	97.8	99.1	100	99.1	99.5	99.1	99.5
MAE	0.24	0.15	0.14	0.12	0.29	0.21	0.16	0.14

Table 3:4. Accuracy ratings for the valence ensemble classifiers (%). Best ensemble version highlighted in grey. The final row also presents each classifier's mean absolute error (MAE) as an overall indicator of its performance.

Error Threshold	Arousal Ensemble Classifiers (3-fold, 10-fold CV)							
	Optimal Classifier Set				Full Classifier Set			
	Averaged Voting		Weighted Voting		Averaged Voting		Weighted Voting	
0.2	77.9	69.0	83.7	71.4	61.5	71.4	76.0	69.0
0.5	93.3	81.0	92.3	83.3	88.5	83.3	90.4	83.3
1.0	97.1	88.1	96.1	88.1	94.2	88.1	96.2	88.1
MAE	0.17	0.38	0.15	0.32	0.29	0.37	0.20	0.33

A Grounded Real-Time Approach

While both the previously discussed approaches were able to classify arousal and valence with a satisfactory degree of precision, they are not adequate for use in real-time applications or most laboratorial studies due to three main reasons:

Requirements for a real-time system are fundamentally different from an offline approach

1. Training cost: Although the developed approach has a low computational cost (both in terms of time and spatial complexity), it may still have a considerable impact in adapting the system to new domains
2. Technically complex: Using AI models implies that most researchers outside the field of AI or non-technical end-users will not know how to properly use said models or train on new datasets
3. Non-continuity: Despite presenting high accuracy ratings, there is no guarantee that when presented with real-time data, the

fusion or ensemble approaches will present a continuous classification output, which, in some cases, is more desirable than high precision (e.g., offline analysis of affective data).

With these issues in mind, we took this chance to develop a simpler version of our system that would address them by requiring a low computational cost, little to no background on artificial intelligence and that presented a continuous output (see Figure 3:5).

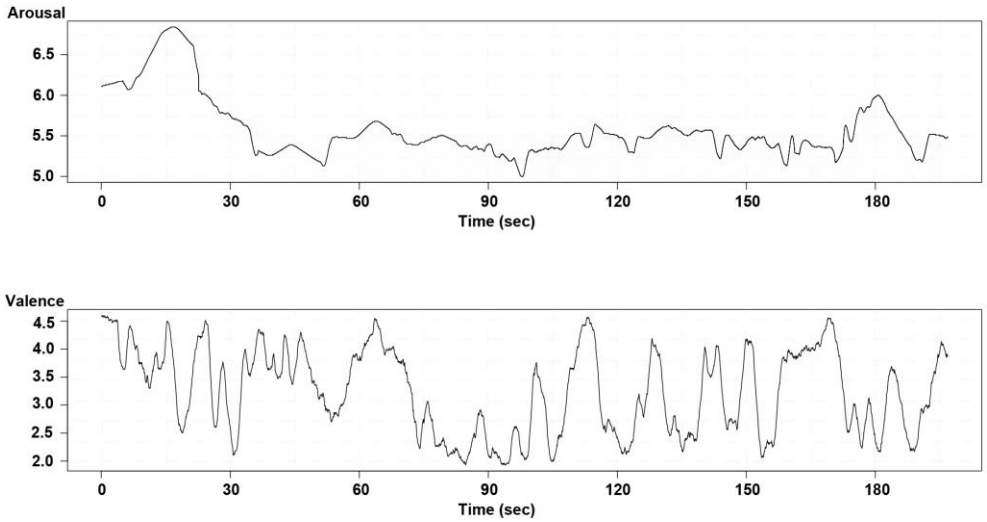


Figure 3:5. Continuous ES estimation for a participant over a 200-second epoch. Analysing the continuous time series may reveal more than simple statistics and thus allow a deeper analysis - even at the cost of lower precision.

For this simplified version, we analysed the literature in physiological computing described in section 3:1 and settled on an approach grounded on the established relations between physiological metrics and the AV dimensions. As in our previous approach, we generated arousal from the regressed SC and HR metrics and valence from the regressed two EMG metrics and also from HR. Although it may sound somewhat simplistic, this type of approach is common in the literature and has shown itself to be reliable and well accepted for most studies. The most prominent of these being the works of (Drachen et al., 2010), (RL Hazlett, 2006) and (Regan Lee Mandryk, 2005), to name a few.

The currently accepted rule of thumb is that SC should be used for assessing arousal, with HR as a fallback or low weight modulator and that valence should be predicted by facial EMG, again with HR as a fallback when no or contradictory facial expressions are being detected - (Regan Lee Mandryk, 2005) being the most popular available example.

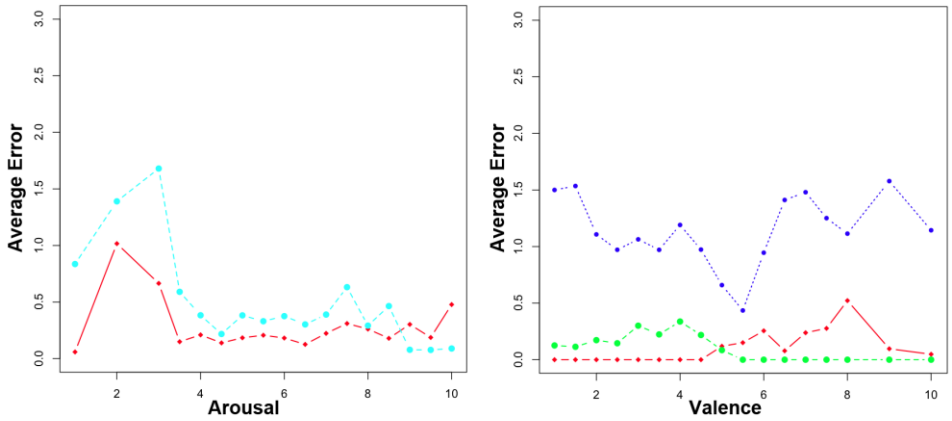


Figure 3:6. Average error values for each of the five regression models created in the first classification layer. Left panel: SC (red) shows a much smaller error than HR (blue) in classifying arousal, except for high arousal values, perhaps due to the HR's more immediate response or saturation issues on some participants. Right panel: HR (blue) exhibits a much higher relative error than both the EMG channels (Zygomaticus in red and Corrugator in green). Notice that each EMG channel was only used to classify either positive or negative valence, hence the symmetrical nil error rates between them.

Using the regression models' error functions as an objective data source for determining our rule system

However, even though this rule of thumb effectively leverages, or rather emphasises, the higher precision that SC and facial EMG metrics have over HR - a phenomenon clearly observed on our regression models (see Figure 3:6) - it is, in our opinion, too generic since there are no conclusive proofs that SC and facial EMG metrics outperform HR every time in respect to arousal and valence prediction, respectively. In light of this, we examined the average error functions of each correlation channel over its predicted AV dimension (see Figure 3:6) and adapted the aforementioned rule(s) to more accurately reflect each metric's real contribution towards an optimal prediction. These adaptations also had the added benefit or stabilizing the method's output, which acted as a makeshift countering mechanism to the variability issues in traditional manual approaches discussed in (R. Mandryk & Atkins, 2007). The classification algorithm created to apply the defined rules can be examined in the appendix section (Appendix A) at the end of this thesis.

Overall, SC was positively correlated with arousal and thus increasing RSC values were mapped to increasing arousal. Likewise regressed HR values also mapped to increased arousal values. Since HR showed a decreasing amount of error for high arousal values, its weight was increased as it predicted higher arousal values. On the other hand, since in other situations it was only allowed to contribute if its prediction didn't distance itself much from the one presented by the regressed SC. Intuitively, HR fluctuates much more than SC and is thus less reliable. However, SC is less reactive than HR and has a slow decay so it may not detect subtle changes. It comes to reason that if SC and

HR predict a similar value, then HR may be more effective in identifying arousal at a more accurate level. Since the EMGzyg channel reflects smiling activity, this channel was correlated with positive valence. Likewise, EMG_{corr} reflects frowning and was correlated with negative valence. Additionally, when both EMG channels were equally active (i.e., predicted near reflexive valence values), the valence output resolved to a neutral value. Finally, since it is possible that an individual shows no facial muscle activity - it is, in fact quite common for most activity to be near neutral (Stern et al., 2001) - HR was used as a fallback measure to the two facial EMG metrics. However, to avoid some of the aforementioned variability issues, we did not allow the classifier to shift suddenly between the facial EMG and HR metrics when the latter became active; instead, HR was given a minimal vote that was always in effect and gradually decreased as either or both EMG channels became more active. For a detailed description of these rules, please refer to the appendix section.

An (unfair) comparison between our grounded and performance-targeted approaches

To maintain the obtained results comparable, the grounded approach was tested on the same testing dataset as the previous approach. Table 3:5 presents the obtained results. These show an expectable loss in predictive power for both arousal and valence. This loss was, most noticeable in the lowest error threshold category, with an average loss in ~25% for arousal and ~30% for valence. However, while arousal quickly recuperates by increasing the error threshold to 0.5 AV units (which in our opinion is an acceptable error for most real-time experiments and simulations), this is not the case for valence. In fact, valence never quite manages to reach the accuracy levels for the optimal ensemble classifier. They are, however, still acceptable for the approach's intended use and an improvement on the discussed literature. Both classifiers' behaviour is easily explained through their mean absolute error (also shown in Table 3:5), which was marginally above 0.18 AV points for arousal and 0.58 points for valence. This justifies why the grounded approach presents such a large boost in accuracy when the error threshold is raised above the 0.2 and 0.5 categories, respectively.

Table 3:5. Accuracy ratings for the grounded and manual approaches (%). Results were obtained by comparing the approach's prediction to the same testing dataset as the ensemble classifier.

Error Threshold	Manual Approach		Grounded Approach	
	Arousal	Valence	Arousal	Valence
0.2	54.2	45	66.5	30.6
0.5	90.3	56.4	96.8	61.2
1.0	98.7	78	98.1	82.3
MAE	0.22	0.58	0.18	0.58

3.5 COMPARING APPROACHES

To validate our previous approaches we wanted to compare their results to a manual approach obtained by following the previously described rules of thumb. As such, for the manual approach we linked HR and SC to arousal and both EMG channels to valence. All channels contributed with equal weight towards the final prediction, so unlike the grounded approach, their error functions were not taken into account. The approach's results can be inspected in Table 3:5. As with the manual approach, the used rules can be examined in the appendix section.

A holistic comparison between all of the developed models and the manual baseline method

For a more complete comparison between our method and the manual approach, we used the differences between their classification accuracies as a distance metric. The obtained results can be observed in Figures 3:7 and 3:8, which present the distance matrices between the manual approach and the remaining ones. Due to space limitations we were not able to present the distance matrices for all error threshold categories. Thus, we chose the ones where results were most significant; 0.2 for arousal and 0.5 for valence.

Arousal	Manual	Grounded	NN	RF	OEA	OWE
Manual	0					
Grounded	12,3	0				
NN	11,7	-0,6	0			
RF	17,3	5	5,6	0		
OEA	17,95	5,65	6,25	0,65	0	
OWE	25,8	13,5	14,1	8,5	7,85	0

Figure 3:7. Distance matrix for arousal classification (0.2 threshold category) between our approaches and the manual approach. Positive values correspond to improvements over the manual approach. Distances are presented in percentage points.

Valence	Manual	Grounded	NN	RF	OEA	OWE
Manual	0					
Grounded	4,8	0				
NN	24,95	20,15	0			
RF	28,4	23,6	3,45	0		
OEA	30,75	25,95	5,8	2,35	0	
OWE	31,4	26,6	6,45	3	0,65	0

Figure 3:8. Distance matrix for valence classification (0.5 threshold category) between our approaches and the manual approach. Positive values correspond to improvements over the manual approach. Distances are presented in percentage points.

Overall, the distance matrix for arousal shows that the grounded approach was able to improve significantly (~12%) on the manual approach, while the best ensemble managed to boost this improvement by a factor of two. Regarding the valence distance matrix, the grounded approach was only able to marginally improve on the manual approach's results – possibly indicating that adding HR measures is not sufficient for accurately describing valence. This could, however be a side effect from the data annotation process, since events tended to elicit clear emotional responses and adding HR measures to identify valence is aimed at improving low energy valence states. Since the highest valence recognition improvements were registered by the RF (28%) and OWE classifier (31%), we believe it is possible that these classifiers were able to capture more subtle aspects relevant to valence classification that eluded both the manual and grounded approaches.

Although these results are encouraging and in our opinion validate our work, there is still clearly room for improvement in what regards valence classification, especially in what pertains to more theoretically-based methods. We further comment on how this might be achieved on section 3:6 (Discussion).

3.6 DISCUSSION

Having presented the results for each of the components of both our approaches, we will now discuss our findings, along with some of the method's conceptual features.

Our first conclusion is related to the observed correlation indexes between recorded physiological metrics and AV dimensions. Although HR did significantly correlate with valence in the relaxing music and

Slenderman conditions, it did not for the IAPS images condition. While this does not impact on the system's performance for our study scenario, it indicates that although it is possible to estimate valence from cardiovascular measures, the same correlation may not always apply and should be confirmed prior to the system's calibration in new scenarios.

Physiologic correlations with emotional states are context-sensitive

However, the main contribution presented by our system's regression modelling phase is that it allows for scaling issues in emotional elicitation ranges and individuals' physiological activation functions, thus correctly unifying the collected data samples.

A detailed analysis of Table 3:2 also reveals some interesting conclusions regarding each of the fusion classifier's performance. Naturally, the best classification accuracies occur when the acceptable error margin is increased. Based on the results presented in the literature, we considered 0.5 to be an acceptable error margin.

Also concerning the fusion classifiers, the highest arousal classification accuracy (~97.5%) is obtained using the single-layer neural networks, closely followed by the SVM classifier (95.6%). Furthermore, there is a clear division between the performance of the NN and SVM classifiers and the DT and RF classifiers. This perhaps hints that the underlying concepts for the former ones are better suited at classifying this type of data. The methods exhibited overall consistent average classification error values, with the lower ones belonging to the RF classifier.

ML classifiers may be accurate, but are also more volatile and difficult to scrutinise

Regarding valence, results are lower than for arousal, but the same general trend applies, with neural networks presenting the highest classification accuracy (~91%). However, the previous divide between the NN / SVM and DT / RF classifiers does not hold. In fact, random forests present themselves as, in our opinion, the *de facto* best choice given their average classification error being the lowest between all classifiers (24% to 57% lower, in comparison).

In sum, we consider that for the tested scenarios, since arousal and valence are distinct concepts, the underlying classification logic may differ and thus, the best choices for AV recognition may be neural networks and random forests, respectively.

Using an ensemble approach to take advantage of these diverse classification logics of our previous classifiers, we were able to further reduce the mean absolute error on arousal classification from 0.28 to 0.13, while maintaining a classification accuracy of 97.4%. We were also able to improve valence classification by reducing the mean absolute error from 0.4 to 0.32 and, at the same time, actually benefiting from a

minimal raise from 91% to 92.3% accuracy. Since the best fusion classifiers tend to be highly volatile and the ensemble method employs a set of different votes towards its final classification, we feel that this approach poses a significant improvement in terms of stability and reliability to our previous results.

However, if scrutiny is not a priority, an ensemble approach can drastically reduce their volatility issues

The most important drawback in employing complex ML models is that they are created in a black-box fashion and are thus difficult to interpret by a human subject, making it hard to evaluate whether the constructed model is overly complex or not. This implies that a direct comparison between the classifiers and traditional theoretical models of emotion requires an indirect, parallel validation approach, which may require multiple validation scenarios and a highly intensive validation workload.

Despite the ensemble classifier's high accuracy, it may not always be an optimal choice due to its computational cost. For example, it might be too costly to execute in some real-time or limited energy systems and its non-continuous output is less than ideal for some tasks. In light of this fact, an approach grounded on the current emotional theories and available empirical data was developed. This grounded approach also represented our attempt at ensuring our method benefited from a full applicability spectrum.

Despite its lower performance, the grounded approach presents itself as a reliable, continuous measure of emotion that surpasses the literature's baseline

Our results show that despite the method's considerable loss in accuracy towards the ensemble classifier, it still managed to attain good results with a larger error threshold. Again, valence presented itself as a much more challenging issue and we believe that either more complex physiological metrics are necessary for achieving high valence classification or these systems should be augmented with contextual information channel to aid the classification process - perhaps a balance of both.

Finally, we wanted to validate all our results. As such, a manual approach based on current good practices described in the literature was developed and its difference in predictive power used as a distance metric. Ultimately, this analysis revealed that all our previous approaches were on the right track, while at the same time able to improve on the manual approach's results.

As it stands, our approach can be used to either accurately identify emotional reactions to specific events over large time windows using a considerably fine-grained scale or in real-time with a continuous output, but with a lesser accuracy. Thus, it represents a contribution not only towards affect detection, but also towards human emotion recognition and thus emotional agent modelling.

3.7 SUMMARY

Emotional detection is not only a critical component of a large majority of affective computing applications, but also a highly complex task with no major consensus by the scientific community and largely lacks (complete) development guidelines.

In this chapter we have presented a theoretical/data-driven hybrid, multi-layered method to interpret selected psychophysiological measures in terms of the arousal and valence affect dimensions. The exhibited results show that we are able to successfully address the recurring emotional scaling and participant physiological activation function normalisation issues present in the literature through a combination of careful experimental design and a mixture of linear and low-degree polynomial regression models. Furthermore, this regression process allows us to keep the system complexity relatively low and humanly interpretable, while also generalising the second classification layer, which means that upon an initial calibration of the regression models, the method is participant-independent.

Since we were able to build our system in a participant-independent fashion, it has shown to adequately generalise within the considered affective experiences and population, although subsequent tests on a larger population are needed for a strong generalisation proof outside these controlled experimental conditions and demographic.

Finally, we were also able to create a simplified version of our method for usage in real-time or low fidelity systems. Our validation tests using a manual approach based on the literature's good practices also showed that both the simplified and the previous approach were on track and managed to improve on its results.

Within the overarching scope of this thesis, this chapter addressed two of our objectives:

Contributions present in this chapter

(II) Define which existing state of the art methodologies are best suited for quantitatively and non-intrusively measure emotional states, while identifying limitations and potential improvements

(III) Develop a generic method capable of measuring the relevant emotional states. The method should provide a continuous measure of emotion in real-time, while also requiring as minimal calibration as possible. Perform a detailed comparison / validation in a relevant (to this thesis) case study with the previously identified methodologies.

Next steps towards creating dynamic, affective user experiences

Having established how to capture players' emotional states in a real-time fashion, we now turn our attention towards creating an emotionally-driven biofeedback game. Doing so will allow us to: *1)* test the effectiveness of emotionally-adaptive games and their impact on players' emotional states (i.e. assess whether they are able to significantly alter players' affect), and *2)* study the correlations between players' emotional states and various dimensions of user experience, such as immersion, tension, and flow, amongst others (thesis objective IV). This study will also double as a data collection phase as we will take the opportunity to collect players' emotional responses to game events so that we are later (Chapters V and VI) able to use them as the basis for their emotional reaction models.

REFERENCES FOR CHAPTER III

Brown, L., Grundlehner, B., & Penders, J. (2011). Towards wireless emotional valence detection from EEG. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society, 2011, 2188–91. doi:10.1109/IEMBS.2011.6090412

Cavazza, M., Pizzi, D., Charles, F., Vogt, T., & André, E. (2009). Emotional input for character-based interactive storytelling. In Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1 (pp. 313–320). International Foundation for Autonomous Agents and Multiagent Systems.

Chanel, G., Kronegg, J., Grandjean, D., & Pun, T. (2006). Emotion Assessment : Arousal Evaluation Using EEG's and Peripheral Physiological Signals. In Proc. Int. Workshop on Multimedia Content Representation, Classification and Security (pp. 530–537). Springer.

Drachen, A., Nacke, L. E., Yannakakis, G., & Pedersen, A. L. (2010). Correlation between Heart Rate, Electrodermal Activity and Player Experience in First-Person Shooter Games. In Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games (pp. 49–54). ACM.

Figueiredo, R., & Paiva, A. (2010). "I want to slay that dragon" - Influencing Choice in Interactive Storytelling. In Digital Interactive Storytelling.

Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004). Emotion recognition using bio-sensors: First steps towards an automatic system. Affective Dialogue Systems.

Hazlett, R. (2006). Measuring Emotional Valence during Interactive Experiences : Boys at Video Game Play. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 1023–1026).

Hazlett, R., & Benedek, J. (2007). Measuring emotional valence to understand the user's experience of software. International Journal of Human-Computer Studies, 65(4), 306–314. doi:10.1016/j.ijhcs.2006.11.005

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system (IAPS).

Leite, I., Pereira, A., Mascarenhas, S., Castellano, G., Martinho, C., Prada, R., & Paiva, A. (2010). Closing the Loop: From Affect Recognition

to Empathic Interaction. In 3rd Int. Workshop on Affect Interaction in Natural Environments.

Leon, E., Clarke, G., Callaghan, V., & Sepulveda, F. (2007). A user-independent real-time emotion recognition system for software agents in domestic environments. Engineering Applications of Artificial Intelligence, 20(3), 337–345. doi:10.1016/j.engappai.2006.06.001

Levillain, F., Orero, J. O., Rifqi, M., & Bouchon-Meunier, B. (2010). Characterizing Player's Experience From Physiological Signals Using Fuzzy Decision Trees. In IEEE Symposium on Computational Intelligence and Games (CIG) (pp. 75–82).

Mandryk, R., & Atkins, M. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human-Computer Studies, 65(4), 329–347. doi:10.1016/j.ijhcs.2006.11.011

Mandryk, R. L. (2005). Modeling User Emotion in Interactive Play Environments: A Fuzzy Physiological Approach.

Moreira, V. H. V. G. (2010). BioStories Geração de Conteúdos Multimédia Dinâmicos Mediante Informação Biométrica da Audiência.

Nacke, L. E. (2013). An introduction to physiological player metrics for evaluating games. In Game Analytics (pp. 585–619). Springer London.

Nasoz, F., Lisetti, C. L., Alvarez, K., & Finkelstein, N. (2003). Emotion Recognition from Physiological Signals for User Modeling of Affect. In Proceedings of the 3rd Workshop on Affective and Attitude User Modelling. Pittsburgh, PA, USA.

Pedersen, C., Togelius, J., & Yannakakis, G. N. (2009). Modeling Player Experience for Content Creation. Computational Intelligence and AI in Games, 2(1), 121–133.

Plutchik, R. (1980). A General Psychoevolutionary Theory of Emotion. Emotion: Theory, Research, and Experience, 1(1), 3–33.

Rokach, L. (2005). Ensemble-based classifiers. Artificial Intelligence Review, 33(1-2), 1–39.

Russel, J. A. (1980). A Circumplex Model of Affect. Journal of Personality and Social Psychology, 39(6), 1161–1178.

S, A., E, W., P, N., J, H., L, M., H, S., & H., K. (2005). Regression analysis utilizing subjective evaluation of emotional experience in PET studies on emotions. Brain Res Brain Res Protoc., 15(3), 142–154.

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). Psychophysiological recording (2nd ed.). New York: Oxford University Press.

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning. Psychological Methods, 4(14), 323–348.

Vinhas, V., Silva, D., Oliveira, E., & Reis, L. (2009). Biometric Emotion Assessment and Feedback in an Immersive Digital Environment. Social Robotics, 307–317.

Yang, Y.-H. (2008). A Regression Approach to Music Emotion Recognition. IEEE Transactions on Audio, Speech, and Language Processing, 16(2), 448–457.

Yannakakis, G. N., & Togelius, J. (2011). Experience-driven Procedural Content Generation. Transactions on Affective Computing, 2(3), 147–161.

Chapter IV

STATIC INDIRECT BIOFEEDBACK

PHYSIOLOGY-DRIVEN MODULATION OF AFFECTIVE PLAYER EXPERIENCES IN A PROCEDURAL HORROR GAME

OUTLINE

In this chapter, we study the impact of static (rule-based) game adaptations based on physiologically-inferred emotional states on player experience. To this end, we developed a procedural horror game (Vanish) capable of run-time level, asset, and event generation.

Vanish was augmented to interpret players' physiological data as a simplified emotional state, mapping it to a set of adaptation rules that modify the player experience. To explore the effects of adaptation mechanisms on player experience, we conducted a mixed-methods study on three different versions of the game, two of which integrated varying biofeedback mechanisms. Players' affective experiences were objectively measured by analysing their physiological data using the grounded approach described in the previous chapter. Additionally, subjective experience was recorded through the use of the Game Experience Questionnaire (GEQ).

Our study confirmed that biofeedback functionality had a statistically significant effect on the ratings of player experience dimensions: Immersion, Tension, Positive Affect, and Negative Affect. Furthermore, participants reported noticeable differences in player experience, favouring the added depth present in the biofeedback-enabled iterations of the game.

The main contribution present in this chapter is the notion that even on its simpler (static) form, biofeedback mechanisms have the potential to significantly change the way players perceive the gameplay experience. It thus lays the basis for the idea that given the ability to learn from players' reactions, these biofeedback mechanisms can become partly intelligent and self-adaptive, leading to an even more lasting and refined human-computer interaction.

As we have previously discussed, video games have drastically evolved over the last two decades, featuring an immense depth of graphical realism and structural world complexity (e.g., virtual, functioning organizations, complex narratives, and even credible social interactions). These have successfully and largely contributed in immersing players as fully as possible. Given this high fidelity level and with the slowing advancements on many of these fronts, the game

research community has begun to focus their efforts on promising and underexplored areas of player experience, seeking to enhance it by using new technologies to increase engagement and motivation.

Chapter objectives and their relation to this thesis' overarching goals

In this chapter, we are interested in enhancing players' experience by introducing biofeedback-enabled game mechanics that use players' physiological data to drive and modify gameplay. By examining the effects of biofeedback mechanisms on player experience, we aim to: *a)* study the impact of static biofeedback mechanics on players' emotional states and gameplay experience, and *b)* evaluate the correlation between players' emotional states and the various dimensions of user experience, thus being able to provide insights for the design of immersive and engaging games. In particular, we focus on the player's affective experience, with an emphasis on their emotional spectra.

In the research community, there has been a great deal of recent interest in affective player experience; various authors have presented studies focusing on affective gaming (Bersak et al., 2001; Dekker & Champion, 2007; Kuikkaniemi, Laitinen, & Turpeinen, 2010; Regan L. Mandryk et al., 2013; L. E. Nacke, Kalyn, Lough, & Mandryk, 2011; A. Parnandi, Son, & Gutierrez-Osuna, 2013a; Avinash Parnandi & Gutierrez-Osuna, 2014; Pope, Stephens, & Gilleade, 2014). Affective games are capable of adapting to the player's emotional state in some way. Currently, a general validation of physiologically-adaptive technology as a successful game mechanic exists. However, the understanding of this technology's full potential is limited (e.g., regarding its emotional elicitation ability or its potential to serve as a regulatory factor for game difficulty or player stress).

Defining relevant gameplay experience metrics

By investigating the ability of biofeedback mechanics to enhance players' gameplay experience, we explore the potential of creating games that are truly immersive and emotionally aware. To accurately obtain an indication of player experience, we use both physiological and subjective measures. Specifically, effects on player experience were measured through the Game Experience Questionnaire with a set of seven dimensions – Immersion, Tension, Challenge, Competence, Flow, Positive Affect, and Negative Affect. While the impact of biofeedback mechanisms has been previously studied in the literature, prior research (e.g., Nacke, Kalyn, Lough, & Mandryk, 2011) has been mostly dedicated to simple mechanics designed as an advantageous player tool (i.e., not on how the game itself might react to players' emotional states). Thus, so far, no study has been conducted to compare the merits of various biofeedback mechanisms in a commercial video game setting. More importantly, the differing effects of these mechanics on the affective player experience have not yet been studied.

The key question progressing logically from this line of inquiry is whether, ultimately, it is possible to modulate players' subjective and/or affective experience using real-time biofeedback game adaptation mechanics. In this case study, we attempt to modulate the presented user experience dimensions in a horror video game. Subsequently, we are also interested in researching the differences between varying implementations of these biofeedback mechanisms and how different players experience the same gaming conditions (i.e., the expectable population variability). These research questions are described in further detail and tied with our thesis' hypotheses in section 4:2.

4.1 RELATED WORK

Reintroducing the concept of real-time physiological data input in digital video games to improve players' cognitive, motivational, emotional and overall engagement has been a recent and active research topic in the Human-Computer Interaction (HCI) field. Since video games are highly immersive (and emotionally-engaging), they pose an exceptional application area on which to test the advantages and drawbacks of biofeedback-based interaction mechanisms.

The origins of biofeedback and its modern resurgence

Biofeedback itself was originally developed for medicinal purposes in the 1970s as a training procedure to overcome medical conditions, such as Attention Deficit Hyperactivity Disorder (ADHD) (Pope et al., 2014). However, in the last decade, it has re-emerged as a viable technology for use in ludic applications. This is not only because of the vast number of interaction possibilities available in video games, but also a result of the low-risk domain presented by them — if a biofeedback mechanism malfunctions or is poorly designed, its potential negative impact is orders of magnitude lower in a game research application than in a medical context.

In fact, biofeedback gaming is currently being explored by some notable game industry giants such as Valve (Ambinder, 2011), Ubisoft through its "O.zen"[4] sensor (formerly known as Innergy) and Sony, which has patented emotional recognition technology with gaming applications (Pigna, 2009) and whose PlayStation 4 DualShock™ controller was designed with imbued skin conductance sensors in mind (Loveridge, 2014). Other, more well-known modern applications are Nintendo's Wii Vitality sensor and Microsoft's new Xbox Kinect 2.0, which possess the capability of optical heart rate detection (Narcisse, 2013). Additionally, several hardware manufacturers are attempting to provide inexpensive physiological input solutions that use brain signals

[4] http://www.experience-ozen.com/

(e.g., Emotiv Epoc[5], Neurosky Mindset[6], OCZ Neural Impulse Actuator[7]) and other physiological measures, such as skin conductance, oximetry, electromyography, respiration rates, and electrocardiography (e.g., BITalino[8]).

Throughout this section, we provide a review of the most current technology and research with regard to biofeedback and biofeedback interaction mechanisms for digital video games. While we have already briefly discussed biofeedback techniques for gaming applications in Chapter I, we now present a much deeper and extensive review of the field, as necessary due to the focus of this chapter's contributions.

We start by introducing the origins of biofeedback and proceed to perform a comparative analysis on the most relevant medical applications of biofeedback techniques. We conclude this review by focusing and presenting a comparative description of biofeedback for affective gaming, thus clearly positioning our contributions within the field.

Biofeedback in Games Research and Affective Computing

Contextualising biofeedback in games research

The academic community commonly refers to biofeedback techniques as affective gaming, for which various definitions have been proposed: the most commonly accepted variant being the one available in Gilleade et al.'s ACE model (Gilleade, Dix, & Allanson, 2005). This model consists of three separate modalities: "*assist me*", "*challenge me*", and "*emote me*" that lay the foundations for how the player's emotional state should be used to adapt the gameplay experience.

However, despite the broad spectrum of concepts addressed by affective gaming, not all biofeedback games are necessarily affective games. For example, a game that uses players' chest volume to modulate aim drift while using a sniper rifle is a biofeedback game, but is not (necessarily) an affective game, because the players' emotions are not a factor in their interaction with the game (despite the emotional effect the biofeedback mechanic may have on the player). Similarly, not all biofeedback games are (or should be) treated as qualitatively equal. For example, take the previous game where players' chest volume was used to modulate aim drift when sniping enemies. Now imagine two versions of the game, one where players know that holding their breath affects aim drift and another one where he does not (assume the modulation is not immediately obvious). It comes to reason that the way

[5] https://emotiv.com/epoc.php
[6] http://neurosky.com/
[7] http://www.bcinet.com/products/
[8] http://www.bitalino.com/

players in which will interact with and experience the game will be quite different. In fact, previous work by Nacke et al. and Negini et al. has studied this, and found significant differences in player experience (reported fun), sensor preference, and player arousal using only simple adaptations of biofeedback game mechanics (L. E. Nacke et al., 2011; Negini, Mandryk, & Stanley, 2014).

Establishing an orthogonal categorisation of biofeedback techniques

Biofeedback game mechanics can be divided into two categories: direct biofeedback and indirect physiological input (Nacke et al., 2011). Here, direct and indirect physiological input refers to attributes of game mechanic activation procedures. For example, in direct physiological input, mechanics are activated through physiologically manipulable signals (e.g., increasing a leg's muscle tension to jump higher), whereas indirect physiological input uses indirectly controlled physiological signals (e.g., increasing heart rate to change the game environmental effects such as weather). However, this categorisation is incomplete, because it disregards the variations in player perception and use of the mechanic (i.e., the learning and adaptation effect that the biofeedback loop provides). A simple example would be the relax-to-win game condition (Ambinder, 2011), where the player must relax in order to achieve a competitive advantage. This is an indirect biofeedback design. However, it assumes that the player has a certain degree of knowledge, neither oblivious nor fully aware, of this adaptation mechanic and reacts accordingly by trying to stay calm or playing the game normally. Should the player not *"correctly"* notice how the game reacts to their physiological state, the game's mechanic breaks down and it becomes unable to convey the desired experience. In the case of a therapeutic game, the effects may be more drastic, with the whole procedure becoming ineffective. Thus, we believe that categorising biofeedback games relies on a two-dimensional space with the following orthogonal dimensions:

- Whether the game mechanic can be activated by controllable physiological signals (direct/indirect physiological input)
- Whether the user should be aware of the resulting adaptation. For this second dimension, we borrow the concept of implicit and explicit biofeedback (Kuikkaniemi et al., 2010).

In our opinion, this variable biofeedback game classification suggests a need for intrinsically different game development guidelines – one for each quadrant. Direct/explicit biofeedback games should exhibit easy to control mechanics and be targeted towards simple gameplay mechanics, while games featuring direct/implicit mechanics should also be simple but with more subtle/secondary adaptations that naturally integrate with players' behaviour during normal (non-biofeedback) game sessions. On the other hand, indirect biofeedback

games can employ more complex data transformation/interpretation. Indirect/explicit variants should strive to adapt noticeable aspects of the gaming experience, contrasting with indirect/implicit variants that should adapt more fundamental "*behind the scenes*" aspects, such as level generation and enemy artificial intelligence (AI). Despite this, both indirect variants need an opaque game adaptation logic that does not allow the player to easily manipulate it – lest he break the game's logic by interfering with the mechanics.

Using our biofeedback game classification as a metric for high-level game design guidelines

Although each quadrant has its own advantages (Kuikkaniemi et al., 2010; Nacke et al., 2011), the crushing majority of biofeedback games have focused on indirect/implicit interaction, mostly by interpreting physiological data in affective terms. In the following sub-section, we discuss the relevant literature in biofeedback, starting with a brief contextualisation on more traditional medical applications and then moving onto modern affective video game applications.

Biofeedback for Medical Applications

Originally, biofeedback was designed to aid in medical therapy by helping patients to overcome medical conditions or to perform patient monitoring/assessment (Blanchard et al., 1996; Bryant, 1991). For example, a music therapy approach is presented by (Dong et al., 2010), where the users' negative emotional states are counterbalanced through music. In a similar approach, (Rocchi, Benocci, Farella, Benini, & Chiari, 2008) presented a system to aid body balance rehabilitation by using simple audio frequencies to indicate correct posture. In related work, (Huang et al., 2005) developed a neural motor rehabilitation biofeedback system for use in a virtual 3D world.

Games provide a more effective medium for therapeutic procedures

Due to biofeedback's easy integration with video games, various serious games have been designed to aid in the treatment of medical conditions. For example, a game was presented which targets the treatment of swallowing dysfunctions (Stepp, Britton, Chang, Merati, & Matsuoka, 2011). Riva et al., (2010) proposed a General Anxiety Disorder treatment that triggers changes in the game world based on the patient's heart rate and skin conductance. A very similar biofeedback game ("Nevermind") for fear management based on players' heart rate readings was also designed (Reynolds, 2013).

Several more ludic approaches have also been taken. For example, *"Brainball"* (Hjelm & Browall, 2000) and Bersak's proposed racing game (Bersak et al., 2001) are relax-to-win indirect biofeedback games that introduce a competitive player-versus-player environment where the most relaxed player has a competitive advantage. While entertaining,

the most interesting aspect of these games is their paradoxical design, because they combine two opposing concepts — relaxation and competitiveness. Naturally, in a competitive environment, players feel pressured to win. In turn, this hinders their ability to relax, which further puts them at a disadvantage and thus, under more pressure. This leads, as we will see further along in this chapter, to a positive feedback cycle where the first player to achieve a competitive advantage tends to have increasingly higher odds of winning. See Table 4:1 for a comparative analysis of clinical biofeedback applications.

Poor biofeedback game design may demotivate players and become ineffective as a therapeutic option

Table 4:1. Review of 10 medical and therapeutic applications of biofeedback techniques.

Reference	SS	BF Type	Adaptations	Treatment	Sensors
(Blanchard et al., 1996)	42	Monitor	Thermal feedback	Elevated BP	BP
(Bryant, 1991)	1	Monitor	Muscle exercise regimen feedback	Swallowing Dysfunctions	EMG
(Dong et al., 2010)	4	IBF	Musical excerpts	Music Therapy	EEG
(Rocchi et al., 2008)	8	IBF	Audio Frequencies	Balance Control	ACC
(Huang et al., 2005)	2	IBF	Musical and Visual Stimuli	Motor Rehabilitation	ACC, PST
(Stepp et al., 2011)	6	DBF	Control virtual fish	Swallowing Dysfunctions	EMG
(G. Riva et al., 2010)	24	IBF	Virtual object placement and properties	General Anxiety Disorder	HR, SC
(Reynolds, 2013)	NA	IBF	Audiovisual stimuli (game events)	Fear / Anxiety Disorders	HR
(Hjelm & Browall, 2000)	NA	IBF	Ball movement / orientation	Relax to win	EEG
(Bersak et al., 2001)	NA	IBF	Car acceleration	Relax to win	SC

Affective Gaming

As previously discussed, there are many definitions of affective gaming, which is generally used as an umbrella term encompassing, among other concepts, biofeedback games. The three heuristics proposed by Gilleade's (2005) model (which serves as the most commonly accepted definition) are: *"assist me", "challenge me",* and *"emote me"*. The first heuristic – *"assist me"* – is meant to guarantee that the game provides hints to the problem in question when it senses that the user's frustration levels increase. The *"challenge me"* heuristic is meant to avoid boring the player, by increasing challenge/threat levels in response to low player engagement. Finally, the *"emote me"* heuristic is the most inclusive, stating only that the player must be emotionally

stimulated according to the game designer's intentions. Although very high-level, this work presents an initial framework that can be applied to the yet-unsolved problem of successful affective game design.

Affective game frameworks

Given that these definitions are necessarily broad, Hudlicka has proposed a set of requirements for creating affective game engines (Hudlicka, 2009):

1. Recognise a wide emotion gamut in real-time
2. Provide mechanisms to respond to affective states
3. Dynamically construct affective user models.

As discussed in Chapter I, we have formalised these into our more structured, conceptual architecture, the Emotion Engine (E^2). As we will see throughout this chapter, Vanish was built around the concepts described in E^2, allowing us to fulfil requirements 1 and 2 and collect the data necessary for requirement 3 in a transparent, non-intrusive manner.

Physiology-Based Affective Gaming

One of the most popular commercial games employing biofeedback techniques was presented in Konami's dating simulator *"Oshiete Your Heart"*, where the player's BVP and SC levels influenced the result of the date. This mimicked earlier commercial games, such as the "*Will Ball Game*", by Charles Wehrenberg, developed circa 1973 for the Apple II. A later title, Tetris 64, released for the Nintendo^{64} platform featured an ear sensor that monitored the player's heart rate, using it to modulate the game's speed. Atari also tested an EMG sensor called Mindlink in the 1980s intended for direct biofeedback – player controlled biofeedback mechanics –, but the prototype never left the lab. These systems were seen as obtrusive and unreliable gimmicks that were easily manipulated, because of their simplicity and cost at the time, and failed to add significant depth to games. However, the recent popularity of affective computing has motivated the resurgence of new and improved affective physiological games (Dekker & Champion, 2007; Kuikkaniemi et al., 2010; L. E. Nacke et al., 2011).

From obscure prototypes to natural interaction, human augmentation controllers

Perhaps one of the best-known examples of (affective) biofeedback games is the study presented in (Dekker & Champion, 2007), where the players' heart rate (HR) and skin conductance (SC) were used to modify a level from the video game Half-Life 2 (Valve, 2004). They introduced biofeedback mechanics that affected the game character's movement and special abilities such as invisibility and ethereal movement (passing through walls), along with post-processing (*shader*) effects, enemy *spawning*, and weapon damage modifiers. While it is not possible to

assess whether the game mechanics were well balanced as a whole and avoided making the game too easy, players reported that it was a highly engaging and personal experience, agreeing that the biofeedback version added substantially to the game's depth. In an attempt to create a more balanced player experience, Kuikkaniemi et al. (2010) created a shooter game where in-game actions such as walking, turning, aiming, gun recoil intensity, and firing rate were based on players' skin conductance and respiration rates.

Another well-known study is the work by (L. E. Nacke et al., 2011) where, as we have previously mentioned, a comparative study between direct (controllable) and indirect (indirectly controllable) physiological input has been presented. In this work, both biofeedback modalities were used to augment interaction with a 2D side-scrolling shooter game that used various physiological sensors (respiration (RSP), skin conductance (SC), heart rate (HR), temperature, gaze and electromyography (EMG)) to alter the control of several gameplay components. Among the modified components were: the avatar's speed and jump power, the enemies' size, weapon range and weather conditions. Additionally, the player's gaze could be used to paralyze enemies for a limited amount of time, which was captured via eye tracking. The authors conclude that the biofeedback conditions were more fun than the non-biofeedback control condition, but players reported preferring the direct biofeedback controls to the indirect ones. This leads us to our previous question on proper game balancing and the significance of players' reported preferences, which we will discuss at the end of this section.

Experimenting with novel gameplay features requires dedicated game balancing and testing

A more simple approach by Rani et al. implements a dynamic difficulty adjustment system for a Pong game based on the user's anxiety state (Rani, Sarkar, & Liu, 2005). This study focused mainly on correctly assessing the player's anxiety state by using various physiological channels. A similar game was proposed by (Parnandi, Son, & Gutierrez-Osuna, 2013) where elements of a car-racing game such as car speed, road visibility and steering jitter were modified based on players' arousal values — extrapolated from skin conductance readings.

Speech recognition (or speech tone interpretation to be more precise) and facial expression analysis are also popular input modalities for affective games. While it is arguable that these should not be regarded as biofeedback games, it is also valid to categorise them as such, because we are still considering physiological features unique to each individual that may or may not be actively controlled by them. One such example is the game Emotional Flowers (Bernhaupt, Boldt, & Mirlacher, 2007). In this game, the player's facial expressions are used to influence the growth rate of a virtual flower at regular intervals

during the day. Another representative example is the interactive storytelling system presented in (Cavazza et al., 2009), where the player interacts with a virtual character through voice commands. In turn, the virtual character interprets both the voice commands and the emotional state of the player, and reacts in a manner congruent with these conditions. A similar game was created by Kim et al. featuring a virtual snail that reacts dynamically to the player's emotional state, which, in addition to voice recognition, is captured using ECG, SC, RSP, and EMG sensors (Kim, Bee, Wagner, & André, 2004).

Experiments with commercial games by the industry giants

Perhaps the industry research most relevant to our work concerns the experiments performed by Mike Ambinder and Valve on some of their most popular games: *"Left 4 Dead"*, *"Alien Swarm"* and *"Portal"* (Ambinder, 2011). The first experiment reported in this work used the player's arousal levels — extrapolated via a SC sensor — to modify the AI director of the co-operative survival shooter game *"Left 4 Dead"*. These modifications pertained to in-game events, such as enemy density, health and weapon item placements, and some architectural elements. In the second experiment, players had a limited time to reach the highest possible game score on *"Alien Swarm"*. However, they were subjected to a *relax-to-win* condition, where their arousal state was connected to an in-game countdown to the end of the game — thus creating the somewhat paradoxical positive feedback loop discussed in section 4:1. Finally, in the third experiment, players played an altered version of the game *"Portal"* where the camera and gun aiming actions were de-synchronized (i.e., instead of the gun always aiming at where the player is looking at, it can be pointed in a different direction, similar to in real life). The camera was controlled using the mouse and the gun's aim through the player's gaze coordinates. Overall, all of the experiments were aimed at identifying the potential advantages and pitfalls of biofeedback game mechanics in various respects (difficulty adjustment, relaxation training, and gameplay augmentation respectively) and found that the relatively simple adaptation performed had a significant impact on player experience — even if some took some time getting used to (some players in the Portal experiment reported dizziness and confusion due to not being used to the control scheme).

On a related note, biofeedback has also been used to modify the behaviour of amusement park rides (Marshall et al., 2011). In this work the user's breathing was used to control the difficulty of the Bucking Bronco ride. An interesting aspect of this particular work was that in one of the conditions the ride's difficulty increased in proportion to respiratory activity — thus, the winning strategy was to hold one's breath in order to reap the perk of a bronco with reduced movement. However, the player has to, eventually, resume breathing – thus

receiving the penalty of a harder ride in proportion to the amount of air taken in. In this work, the breath control mechanic provides game balancing since, similarly to the relax-to-win games previously discussed, the game has the ability to polarize the effects of obeying the relaxation mechanic (i.e., gasping for air leads to a higher bronco ride intensity, which then causes players to lose breath more quickly).

Benefits and open issues in biofeedback gaming

Throughout this section, we have examined several studies on the impact and design of biofeedback games (see Table 4:2 for a comparative analysis of the discussed literature). In all of these studies, authors have reported increased engagement and enjoyment factors using various similar metrics. However, some issues remain with the presented analysis, one of the most common being the effective game balancing, which has arisen repeatedly throughout this section. Virtually all of the discussed mechanics were designed as aids to the player, which — despite adding to the gaming experience — may introduce a bias effect on the obtained positive feedback, because they are seen as beneficial to the player. Not all of the discussed work employs this type of mechanic solely, but its dominance remains a relevant discussion point on the impartiality of the reported results. Furthermore, the evaluation of players' preferences is occasionally shallow, using direct questions instead of a more structured and robust questionnaire, or failing to provide a statistical analysis of the obtained feedback to check for statistical significance.

However, in our opinion, the most persistent issue is the lack of an analysis on how dissimilar biofeedback mechanics impact players' subjective gaming experience and also the dismissal of analyses regarding the variance of players' affective (physiological) states during the gaming sessions. Coupled with the absence of any detailed analysis on how different player sub-groups (e.g., *casual* vs. *hardcore* gamers) experience each gaming condition, these issues constitute what we consider to be the most pressing research questions yet to be answered in the current literature.

In summary, the current state of the art has widely proven that biofeedback is commercially viable and has the potential to add to player experience. Conversely, research has yet to thoroughly examine to what degree and in what manner the design of the biofeedback mechanisms themselves can be designed to influence players in an *a priori* defined way.

Table 4:2. Review of 12 biofeedback-enabled affective games.

Reference	SS	BF Type	Adaptations	Game Genre	Sensors
(Dekker & Champion, 2007)	33	IBF	Movement, hearing, audiovisual effects, damage modifiers, game difficulty	FPS (Half-Life 2)	HR, SC
(L. E. Nacke et al., 2011)	10	IBF/DBF	Movement speed, jump power, enemy size, weapon range, weather conditions	2D side-scrolling shooter	RSP, GSR, EKG, TMP, Eye Gaze
(Kuikkaniemi et al., 2010)	36	DBF/IBF	Movement, aiming, gun recoil, firing rate	FPS	SC, RSP
(Rani et al., 2005)	15	IBF	Dynamic difficulty adjustment	Pong clone	ECG, ICG, PPG, Heart Sound, GSR, EMG
(A. Parnandi et al., 2013b)	20	IBF	Car speed, road visibility, steering jitter	Racing game	EDA
(Bernhaupt et al., 2007)	21	IBF	Flower growth	Casual meta-game	Facial expressions
(Cavazza et al., 2009)	14	IBF	Virtual character reactions / narrative adaptation	Drama narrative	Voice
(Kim et al., 2004)	4	IBF	Virtual snail reactions	Casual meta-game	Voice, ECG, SC, RSP, EMG
(Ambinder, 2011)	NA	IBF	Enemy density, item placement	FPS (Left4Dead)	SC
	NA	IBF/DBF	Timer countdown rate	TPS (Alien Swarm)	RSP
	NA	DBF	Aiming mechanism	FPS (Portal)	Gaze
(Marshall et al., 2011)	14	IBF	Ride movements	Amusement ride	RSP

4.2 GOALS & RESEARCH QUESTIONS

As we have discussed in the previous section, our aim is to perform a comparative analysis on how different types of biofeedback gameplay mechanics affect players' gaming experience. Our main objective is thus to thoroughly examine players' subjective (GEQ reported) and objective (physiological) gaming experience, in order to assess whether the design of the aforementioned biofeedback mechanics have significant effects. In parallel and in line with our intended detailed examination, we are also

interested in evaluating whether specific types of players react in a particular way to any of the game variants.

To summarise prior statements, the underlying question is whether, ultimately, by employing specifically designed biofeedback mechanisms, it is possible to modulate players' subjective and/or affective experience — in our case, within the context of a horror videogame.

Research Questions

The question we specifically wanted to address with this study was the one in thesis objective IV: Whether indirect biofeedback gameplay adaptations significantly impacted player experience and, if so, how exactly – i.e. what are the present correlations between player experience and their emotional states?

To approach this question in a more structured manner, it was segmented into the following sub-questions:

Structuring and extending one of our most central research questions

- Q1: Can we modulate the "scary" experience of playing a horror game using physiological sensors and real-time processing, effectively biasing players' emotional states to a predetermined set of target emotional states?
- Q2: Do these biofeedback-enabled adaptive methods have a significant impact on any specific aspect of the players' gameplay experience?
- Q3: How do implicit and explicit indirect biofeedback mechanics compare to each other in terms of user preferences and experience?
- Q4: Do different types of players, as distinguished by sex, proficiency and game genre preference, present any noticeable differences in how they experience these modifications, either in terms of reported player experience or physiologically measured emotional states?

Regarding question Q2, we assume a directed hypothesis: biofeedback conditions increase the ratings and have a stronger impact on player experience, both subjective and objective, than non-biofeedback conditions. Regarding questions Q1, Q3, and Q4, we do not assume directed hypotheses, only that the conditions may differ between them.

Answering these research questions required us to measure several different aspects of the gaming experience. Question 1 required us to have an objective measure of players' affective experience throughout

the gaming sessions. To do this, we provided the game with logging capabilities so that it would output players' raw physiological and emotional state data. As a secondary measure, we also used the Game Experience Questionnaire's positive and negative affect dimensions. To measure players' gameplay experience in Q2, we used the Game Experience Questionnaire, which measured experience along seven distinct dimensions — Immersion, Tension, Challenge, Competence, Flow (Csíkszentmihályi, 2008), Positive Affect and Negative Affect — via a structured set of game-related statements. Additionally, overall Fun was also measured on a similar 7-point Likert scale. Regarding Q3, since it addressed player preferences and reported experience between biofeedback conditions, we chose to ask players to order conditions based on their preferences (while maintaining condition order blinded to avoid bias effects) and also posed some open-ended questions to gather feedback. To answer Q4, demographic information regarding the participants was collected during the recruiting process.

Experimental Constraints

Given that we aimed to evaluate a somewhat broad set of hypotheses, there are several experimental constraints on our study. The following constraints were identified both in our study and game design:

Making sure the collected data and experimental protocol are sound/feasible

(III) *Number of gaming conditions:* Given that we wanted to compare two biofeedback conditions, our study should require three gaming conditions (two biofeedback-enabled and one control). Furthermore, this implied that all participants had to play all three gaming conditions, which required an extensive experimental protocol.

(IV) *Real-time game adaptation:* Implementing two different styles of biofeedback mechanics implied that significant portions of the game would be readily and swiftly adaptable, to react to players' physiological alterations. This meant that the game had to feature a versatile system capable of modifying the following elements in real time: enemy AI, level and asset/event generation, character attributes (e.g., running speed, animations), sound and visual effects, light sources, and item placement.

(V) *Game architecture:* Since we wanted the game to work in a standalone fashion without the physiological data (for the control condition), this required an overall modular architecture (i.e. the used components of the E^2 architecture could be toggled on/off as needed). In this architecture, the emotional detection module could be "plugged into" the

system. In the absence of its input, a *fallback* system would take over to modulate the game experience based on a set of rules that knew nothing of the player's current physiological state; a game AI director, typically found in commercial games.

(VI) *Data logging capabilities:* Being able to properly assess players' physiological data during the gameplay sessions required that the game be able to output this data, trimming it to the intervals, where players were actually participating in the game.

4.3 STUDY

To evaluate the relative impact of different types of Indirect Biofeedback (IBF) adaptation mechanics, we conducted a mixed-methods study where participants played three versions of the game; two augmented using the IBF mechanics and one non-biofeedback control condition. To maintain comparable results, each game version presented the same gameplay elements (assets, AI, events, etc.) and game mechanics. Both the game design and IBF adaptation mechanics were developed during an extended alpha-testing period using an iterative prototype over 3 months prior to this study, gathering feedback from over 20 individuals not included in this study.

Gaming Conditions

Each of the IBF game conditions was designed to map changes in the players' arousal, A, and valence, V, ratings — measured as their partial derivatives dA/dt and dV/dt respectively — to specific adaptations in the game's controllable elements (see section 4:4). The control condition used no physiological input (although players were still wearing the sensor equipment, so they were unable to easily identify under which conditions they were playing).

Linking the player to the game world through his avatar psyche

The first biofeedback game condition aimed to deepen the empathic bond between the player and the game character by binding the emotional states of the player and the game character. To convey this *"symbiotic"* link, the character's emotional state is externalized through its anaerobic metabolism attributes A and mental resilience r. Regarding anaerobic metabolism attributes, increases in the player's arousal ratings mapped to increases in the avatar's adrenaline levels, which translated into higher running speeds a_v, albeit at the cost of stamina a_s (i.e., running time). Conversely, if the player remained calm, he would only be able to run at a more conservative pace, but for much

longer distances. Simultaneously, the avatar's mental resilience, *r*, was directly correlated to the player's valence levels, which translated into more intense hallucinations and uncontrollable camera twitches that impaired the player's sense of orientation and also made them a more easily detectable target for the creature (i.e., the enemy). Since stealthy, paced exploration is the optimal game strategy — (making noise quickly attracts the creature, and running, in addition to making noise, opens the character to being easily blindsided) — remaining calm provided a competitive advantage. We named this IBF condition "Symbiotic IBF". In our orthogonal classification of biofeedback games, this condition is an Explicit Indirect biofeedback game, because the player has no direct control over the physiological signal under analysis and the gameplay mechanics are made evident through their externalization on the game character's own physical and mental attributes. The condition can also be contextualized in terms of Gilleade's '*emote me*' heuristic as it plays directly on the emotional interplay between the player and their avatar.

Linking the very nature of the game world to the player's psyche instead

In the second condition, we aimed to stabilize the player's emotional levels during gameplay by using a dynamic difficulty adjustment (DDA) scheme that altered the level's generation parameters. In this condition, the player's arousal level was inversely correlated with the probability of generating a creature encounter, *c*. However, our alpha tests revealed that, regardless of player arousal, environmental events still needed to occur sporadically to maintain player immersion. Thus, the probability of each environment event e_i occurring presented a direct correlation to the player's arousal level. Subsequent alpha testing also suggested that desynchronizing creature events and environmental events could create increased tension levels, reflected in lower valence ratings. Thus, valence was used to determine the player's level progression rate ρ, meaning that valence was inversely correlated with the generation of rooms associated with goal completion. Additionally, valence was also inversely correlated with the probability p that each newly generated level block b_i would feature an escape tunnel in its configuration, $p(b_i^{tunnel})$. This meant that a player with a high valence rating (perhaps feeling confident in his abilities), was offered a reduced number of possible escape routes for use in subsequent creature encounter events. This contrasts with the previous condition, and is situated in the Implicit/Indirect quadrant of our biofeedback game classification space, because in this second IBF condition the gameplay adaptations are intended to be imperceptible to the player. Also in contrast with the Symbiotic IBF condition, it falls more in line with Gilleade's '*challenge me*' heuristic, because it strives to maintain a balance in the player's emotional states. We thus termed this IBF as the "Equilibria IBF".

Mapping between the player's emotional state dimensions and the gameplay adaptation mechanics for each IBF condition is illustrated in Table 4:3. Notice that, despite the gameplay adaptations occurring whenever an emotional change was triggered, this does not imply an immediate change in the actual game world. It merely alters the game's procedural mechanism parameters (refer to section 4:4 for a description on how these were reflected in the game world).

Table 4:3. IBF Adaptation Mechanisms (see game mechanics description in section 4:4). ε equals 0 and N(x) represents a normalisation function.

IBF CONDITION	Emotional Trigger	Gameplay Adaptation
Symbiotic IBF	$\lvert dA/dt \rvert > \varepsilon$	$a_v \mathrel{+}= N(dA/dt)$ $a_s \mathrel{+}= \min(0, -N(dA/dt))$ $E_{heartbeat} \mathrel{+}= N(dA/dt)$
	$A \geq 9.5$	$E_{faint} = 1.0$
	$\lvert dV/dt \rvert > \varepsilon$	$E_{FoV} \mathrel{+}= \min(0, -N(dV/dt))$
	$\lvert dA/dt \rvert > \varepsilon$ OR $\lvert dV/dt \rvert > \varepsilon$	$r \mathrel{+}= N(dV/dt) + \min(0, -N(dA/dt))$
Equilibria IBF	$\lvert dA/dt \rvert > \varepsilon$	$c \mathrel{+}= \min(0, -N(dA/dt))$ $E_{any} \mathrel{+}= N(dA/dt)$
	$\lvert dV/dt \rvert > \varepsilon$	$\rho \mathrel{+}= \min(0, -N(dV/dt))$
	$\lvert dV/dt \rvert > \varepsilon$ AND *creature.chasing( ) == TRUE*	$\left(b_i^{tunnel}\right) \mathrel{+}= \min\left(0, -N(dV/dt)\right)$

Making the game engine understand and act on these "links"

Experimental Protocol

The study used a three-condition (two IBF-adapted and one control) within-subjects design. After being provided a brief description of the experiment (IBF adaptation details were kept from the participants) and providing informed consent, players completed a demographics questionnaire. They were then fitted with physiological sensors and underwent a calibration phase (described in the following paragraph) designed to correctly tune the emotional detection system. After a brief tutorial, all participants played the three game conditions, which were presented in Latin-Square balanced arrangement to avoid order effects. Additionally, participants were given a resting period of 5 minutes between gameplay sessions during which they were allowed to relax. Their physiological baseline levels were also adjusted, if necessary. After completing each gaming condition (~10-25 min.), participants completed a game experience questionnaire (GEQ) (Ijsseltein, Poels, & De Kort, 2008), which asked them to rate their experience on several dimensions. Additionally, they were also asked to report their fun ratings. To cancel external noise, all participants were left alone in the room and were required to wear a pair of noise-cancelling headphones during the calibration and gaming conditions. Ambient lighting was

Setting up a control condition, eliminating order effects and reducing experimental noise

also controlled (turned off) in order to create a familiar gaming environment and potentiate heightened emotional states.

Sensor Calibration

Since the two biofeedback-enabled conditions required a continuous, real-time emotional state feed to be computed from the player's physiological ratings, it was necessary to calibrate the emotional detection system prior to performing the experiment. Thus, each participant underwent a simple calibration process including the following stimuli:

Given the low elicitation capabilities of the IAPS library, the sensor calibration phase was adapted to use film clips

- Relaxing Music: The participant was asked to put on a pair of noise-cancelling headphones and relax. This phase lasted for approximately 5 minutes.
- Scaring Image Game: The participant was asked to identify a non-existent object on an image. After 20 seconds, the image suddenly switched to a disturbing one, combined with a loud screaming sound.
- Film Clips: Based on the work of (Schaefer, Nils, Sanchez, & Philippot, 2010), we chose two movie clips, one from a comedy film (American Pie: The Wedding) and one from a thriller film (American History X) , to be viewed by participants.

The first two stimuli aimed to elicit low and high arousal values respectively, and the set of movie clips attempted to elicit divergent valence levels. Participants were left alone in the room (similar to the gaming conditions), so that our presence would not influence their emotional states. After experiencing all of the stimuli, we analysed the captured data and asked the participant to rate the AV values that each stimuli elicited, which we then used to calibrate the emotional detection module. To make sure that the emotion recognition module worked as expected during the gaming sessions, we also manually loaded the computed regression coefficients for each physiological metric/AV space dimension, validating its output just before players begun each of the three gaming conditions.

Given that Vanish is the same game genre as Slenderman (the game used in Chapter III for data collection), confirming that our previous correlation assumptions held was arguably a (still necessary) formality. All assumptions were verified for all participants; skin conductance and heart rate positively correlated with arousal, and facial electromyography at the cheek and brow muscles correlated positively and negatively with valence, respectively.

Apparatus

The game was played on a desktop computer running Windows 7. The monitor was a 23" LCD display running at a resolution of 1920x1200 pixels. Physiological data was collected using the Nexus-10 hardware by Mind Media and integrated into the game through the emotional detection system in real-time at 32Hz. Furthermore, SC was measured at the subject's index and middle fingers using two Ag/AgCl surface sensors attached to two Velcro straps. HR was derived from BVP readings measured at the thumb using a clip-on sensor. Facial EMG was measured at the *zygomaticus major* (cheek) and the *corrugator supercilii* (brow) muscles.

Participants

Data was recorded from 24 participants (16 male), aged 19 to 28 (μ=22.5, σ=2.5). Participants were recruited from a pool of volunteers that responded to an online beta-testing recruitment campaign. Unfortunately, because of sensor malfunction during one of the game conditions, the data from one of the participants was corrupted and could not be used. For the remaining 23 participants, no malfunctions occurred.

Players appeared to be rather well balanced in terms of gaming proficiency and game genre preference.

All of the participants played video games at least weekly. Regarding previous experience with horror video games, participants were reasonably equally segmented, with 67% of participants having played a horror game at least once. However, of these 67%, only 47% actually enjoyed the experience. Most participants (58%) rated themselves as *casual* players (less than 4 hours per week of gameplay time), as opposed to the 42% who considered themselves to be *hardcore* players. Most players reported experience and/or interest on novel forms of input for computer games, with a large part of them (64%) owning a Nintendo Wii or mobile gaming platform.

4.4 VANISH: A PROCEDURAL AFFECTIVE HORROR GAME

Vanish (3DrunkMen, 2013) is a free[9] indie survival horror video game built on top of the Unity game engine (see Figure 4:1). We chose a horror video game due to the fact that emotional reactions have an especially strong influence on tense and scary gameplay experiences.

[9] Vanish is available for free on IndieDB (http://www.indiedb.com/games/vanish) for both Mac and PC. After its alpha release on July 2014, it was ranked #8 out of a total of 17,754 games.

This feature is used as a selling point in horror games (e.g., Amnesia – The Dark Descent (Frictional Games)).

Vanish's general game design and premise are that of loss of control and fear of the unknown

In Vanish, the player must navigate a network of procedurally generated maze-like tunnels and odd machinery rooms to locate a set of key items before being allowed to escape. However, the game's main premise revolves around the psychologically-straining experience of being hunted while navigating these tunnels, as the player must evade a deformed creature that continuously — and increasingly more aggressively — stalks him. Other environmental events, such as lights failing, distant cries, or steam pipes bursting without notice, occur semi-randomly to keep the player engaged.

Both the game's level layout and events can be generated at runtime and simultaneously linked to form seamless level sections. This allows us complete control over the game's logic, level generation, AI, and general level progression. As we have discussed in section 4:1, indirect biofeedback games should employ opaque adaptation logic so as to not allow the player to easily manipulate the game's logic. Thus, the (game design) freedom that Vanish conferred us in this regard was a critical factor in its choice.

Figure 4:1. Two game screenshots featuring the player's reactions[10] of two game locations: a machinery room (left) and a creature encounter on a darkened corridor (right).

In the following sub-section, we will contextualize Vanish's most relevant gameplay mechanics, and offer a brief explanation of the game's procedural generation. We conclude by presenting some light technical details on how we adapted the game's architecture to enable the biofeedback mechanics.

[10] All rights for the game session footage belong to PewDiePie: www.youtube.com/watch?v=2FQmdBGomCo

Gameplay Mechanics

To explore the impact of the variations between traditional game mechanics and physiologically adapted ones, we implemented four game mechanics that could be influenced using the player's emotional states.

Character Sanity

The game character has a sanity profile. Over time, the character will become more agitated and his mental state will degrade. This effect is, naturally, greatly accelerated by (negative) game events, such as encountering the creature or being left in the dark. Successfully advancing in the game improves the character's sanity. To perceive this character behaviour, some game mechanics were altered. Our sanity system has the following levels:

- Level 1 — Sane: This is the initial sanity level. No visible alterations to the character's psyche occur.
- Level 2 — Scared: When this level is reached, the character's breathing will become noticeably faster and heavier, with occasional shivers.
- Level 3 — Terrified: The character starts hallucinating and hearing strange sounds behind him.
- Level 4 — Insane: When the character reaches this level he becomes dizzy and the player will have a hard time controlling the camera movement. The hallucinations from the previous level worsen considerably (both in intensity and frequency), and include intense imagery such as insect swarms.

Designing the gameplay mechanics with biofeedback in mind allowed a more natural fit and greater expressivity

Additionally, in the control (non-biofeedback) version of the game, whenever a game event takes place, a set of special effects will take place in accordance with the intensity of the event. This was done to more heavily stimulate and convey the importance of each event to the player. These include a vignette-style filter, which warps the player's field of vision (FOV), a loud heartbeat sound effect, and a camera-shake effect. As previously mentioned, in the biofeedback versions of the game, we also alter these effects to occur in response to changes in the player's emotional state (see Section 4:3).

Creature AI

The creature shifts between three predetermined states: Passive, Passive-Aggressive, and Aggressive. In each of these states, the creature takes different approaches to the player; from fleeing from them, to only approaching if the player gets close, to finally hunting the

This interplay incites a more personal relationship with the creature that intrigues and scares players

player down. A dynamic transition scheme was implemented that depended not only on the players' actions towards the creature, but also on how much interaction there is and how near the player is to finishing the level. If, for example, the player seeks out the creature, it will at first retreat. However, over time, it will start to get "curious" and approach the player more proactively, until the player flees from it. This, in turn, creates a push-and-pull power relation between the player and the creature, which, if well managed, can benefit the player's progression.

Character Sprint Velocity, Stamina and Orientation

Not being able to fight back creates a more emotionally charged and volatile experience

Like other psychological horror games (e.g., Amnesia) Vanish does not provide the player with any means to fight off his aggressors, leaving the player with no options aside from exploration, hiding, and fleeing. This lends the game a tense atmosphere and heightened sense of danger, as well as transforming the relatively straightforward character control functions into central gameplay mechanics. Additionally, with the modifiers acting on the orientation function provided by the sanity and fear mechanics[11], we are able to alter the character's sprint velocity and stamina (measured in the distance he is able to run continuously).

Evasion Tunnels

Tunnels give players a sense of hope and motivation to explore since mistakes are not as costly

As we mentioned in the previous paragraph, cautious exploration is a central aspect to this genre's gameplay. However, no matter how cautious the player, he will eventually come face to face with the creature multiple times. In these situations, it is important to react quickly, flee the site, keep calm, and devise a plan to elude the creature (now in pursuit of the player). To reward quick-thinking players and create dramatic tension events (such as being cornered by the creature and still surviving), we semi-randomly spawn tunnels in strategic positions in the walls along the player's path. The player can crouch into these to escape the creature, waiting for it to go away or exploring them to reach a new level section altogether.

Sound Sources

Since sound is one of our most acute senses and a main dramatic enhancer in horror games, we wanted to incorporate it as a gameplay

[11] While we did not have the required hardware, the game allows the usage of an Oculus Rift device for orientation and omnidirectional treadmill for movement. We find the idea of analysing the effects of modulating movement speed and orientation when using these devices particularly interesting.

mechanic, thus motivating players to pay close attention to the game's atmosphere. This was achieved by making the creature AI reactive to sound — i.e., when within a certain radius of the player, the blind creature will detect him if he makes too much noise (e.g., breathes too heavily or moves around too quickly). This created some interesting scenarios where players were trapped in a dead-end, but by remaining perfectly still, were able to evade the creature; or scenarios where they would normally not attract the creature's attention, but because of a low sanity level, were unable to properly control the character's movements and were killed.

Making players focus on their movement breeds a greater sense of presence, which benefits immersion

Procedural Generation

As we have previously discussed, besides being able to tune the gameplay mechanics' behaviour, Vanish also allows us to control various other aspects of runtime generation with respect to level layouts, assets, and events. Since one of our driving research questions was whether we could modulate the "scary" experience of playing a horror game in a semi-automatic fashion (i.e., without resorting to scripted sequences that rapidly become predictable, and thus boring), we decided to adopt a *"Game AI Director"* approach.

The term Game AI[12] Director, or simply AI Director (AID), refers to a dynamic system that continuously considers the player's progression and behaviour. Based on these, it then changes the gameplay according to a specific rule set or internal logic — in our case, these were the game's original gameplay mechanics (described in this section) and biofeedback adaptation mechanisms (described in Section 4:3). Popular games using this type of system include, for example, the Binding of Isaac (Headup Games, 2011), Left4Dead 2 (Valve, 2009) and Driver: San Francisco (Ubisoft, 2011), among others.

Procedural generation as the main pillar in the game's adaptability

So as to allow the AID to perform necessary adjustments to the game levels, we combined it with a procedural content generator that determines how and when to generate the required content. However, while the content generator is capable of generating virtually any available game element, Vanish imposes an underlying structure dictating the manner in which levels should be constructed. To represent this underlying structure, we used a design grammar framework (DG) to describe valid level configurations (see Figure 4:4).

[12] Game AI is a term referring to a specific subtype of AI usually used in games. It usually refers to techniques that create intelligent-like behaviours on behalf of Non-Player Characters (NPCs), such as player interactions, strategy formulating or pathfinding. Sometimes it extends to other areas of the game world, such as terrain or level generation in procedural content generation or player experience or behaviour modelling. A Game AI Director is a type of system that oversees the AI of most aspects of the game, including assigning group strategies to several NPCs to simulate coordination (Bourg, 2004).

A chunk-based approach also helps alleviate potential content explosion issues

Like in other procedurally generated games (e.g., Binding of Isaac, Minecraft), Vanish uses a *chunk*-based approach for generating levels – i.e., game levels are created by arranging different chunks (i.e., level "pieces") in a three-dimensional space. To allow organic change of the map during playtime (see Figure 4:2), chunks outside of a set radius around the player character are automatically *despawned*. Several adaptations to the game AI's *pathfinding* and game logic were made to accommodate this feature, but they remain out of the scope of this thesis.

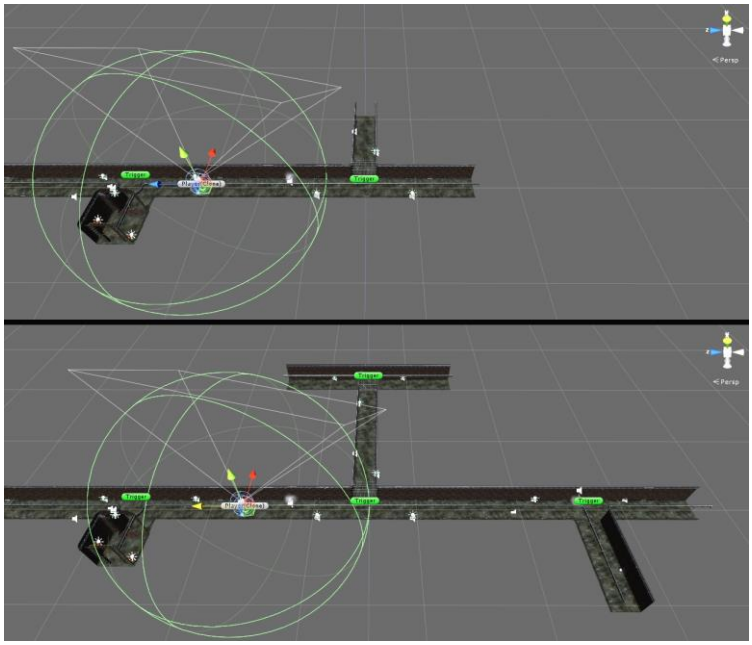


Figure 4:2. Unity scene view portraying the spherical spawning mechanism. Top: Moment before the sphere enters a new block. Bottom: New blocks spawned in response to the triggering mechanism. Blocks are seamed together by using anchor points in key locations of the block's base template.

Chunks must be customizable to avoid too much repetition

Similarly, each chunk is created from a *"clean"* base template that contains multiple anchor points used to spawn the creature or game elements such as lights, collectible items, or triggerable game events (see Figure 4:3). This simplifies the level generation process, while also making it more versatile.

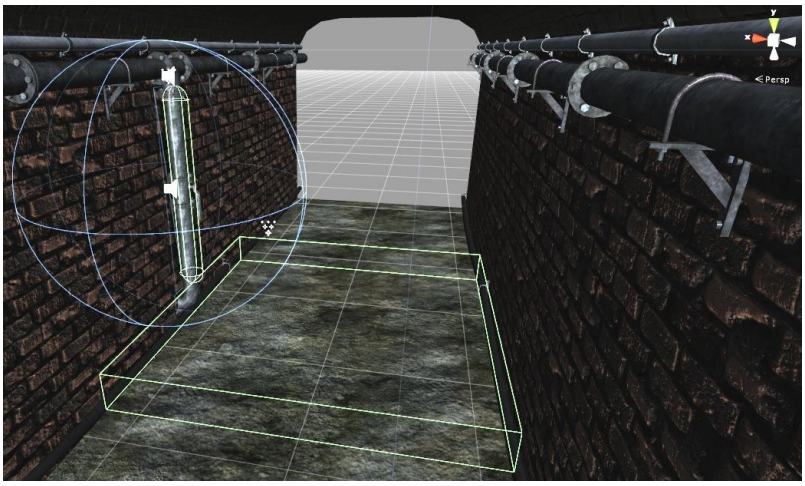


Figure 4:3. Unity's game development view. The green box near the floor represents the collider responsible for triggering, with a certain probability p, the "*pipe burst"* event once the player steps in it.

Representing valid game world configurations as efficiently and compactly as possibly

In our design grammar, a level is specified as a set of blocks and a creature (which can either reside within the map, or not, depending on its spawned state). Each block (chunk) can be a type of connecting corridor, a dead end, an exit, a machinery room, or the creature's nest. Each block also has an x and y coordinate, as well as an orientation and a list of linked assets (its configuration) that can include items, a tunnel escape, or a set of game events. A simplified version of the DG is presented in Figure 4:4 and details all of the possible blocks' types, events, and parameterization possibilities. Additionally, level configurations that "boxed-in" the player (e.g., two dead ends connected to each other) were considered invalid.

While the DG provides the generator with valid potential level configurations at runtime, the generator is also responsible for determining which events and level progression should take place at any given time. To achieve this, the generator is capable of adjusting each of the following controllable game elements:

- **Block Type Frequency:** The frequency of each level block type λ_i such that $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, for n types of level blocks.
- **Level Progression:** The level's progression rate ρ per traversed world block (i.e., the player's advancement towards a set of sequential goals), expressed as the average of the game reward (blocks featuring an objective item) frequency rates: $\rho = {}^{1}/_{\sum_1^n E(r_i)}$, where $E(r_i)$ denotes the expected value of a reward r_i that occurs anywhere between [0, n] block generation turns.

```
<level> ::= <blocks> <creature>
<blocks> ::= <block> | <block> <blocks>
<block> ::= oneway_corridor(<x>,<y>, <orientation>, <assets>)
  | threeway_corridor(<x>,<y>, <orientation>, <assets>)
  | fourway_corridor(<x>,<y>, <orientation>, <assets>)
  | dead_end(<x>,<y>, <orientation>, <assets>)
  | exit(<x>,<y>, <orientation>, <assets>)
  | machine_room(<x>,<y>, <orientation>, <assets>)
  | creature_nest(<x>,<y>, <orientation>, <assets>)

<assets> ::= <asset> | <asset> <assets>

<asset> ::= <event_trigger> | <item> | <tunnel>

<tunnel> ::= (<spawn_reference>, <orientation>)

<event_trigger> ::= (light_burst | waterpipe_burst |
      steampipe_burst | pipe_fall | hallucinations |
    environment_sound | creature_sound | explosion |
    camera_distortion | player_faint) <triggering_prob>
    <spawn_reference>

<item> ::=  (paper_note | folder | light_sticks) <spawn_reference>

<creature> ::= <AI_mode> <spawn_reference> <orientation>

<AI_mode> ::= (idle | patrol | chasing | searching)
<spawn_reference> ::= [-1..10]
<orientation> ::= (north | east | west | south | northeast | north-
west | southeast | southwest) <offset>

<offset> ::= [-45..45]
```

Figure 4:4. A simplified version of the game's design grammar. The rules expressed by this BNF grammar are internally used by the game's logic to, at each content generation cycle, assert whether the new level blocks and sections form a valid configuration, thus avoiding unplayable or broken levels.

Formalising the controllable game elements

- **Mental Resilience:** The game character's mental resilience $\{r \mid r \in \mathbb{Q}^+ \wedge 0 < r \leq 2\}$, which acts as a base modifier for the standard event values for Sanity and Immediate Fear gameplay mechanics.

- **Anaerobic Metabolism Attributes:** The game character's anaerobic metabolism (the rate of the body's metabolic energy expenditure on high intensity activities) attributes, $A = (a_v, a_s): a_v, a_s \in \mathbb{N} \wedge a_v, a_s \geq 0$, where a_v represents the character's running velocity and a_s its maximum stamina (i.e., for how long the character can sprint uninterruptedly).

- **Creature Encounters:** The frequency of encounters with the game's creature$\{c \mid c \in \mathbb{Q}^+ \wedge 0 \leq c \leq 1\}$, measured in terms of its spawning probability (on a spawning point of a random level block not within the player's FOV) at each newly generated level block.

- **Environment Events:** Given the n possible environment events E in a game level L comprised of a set of k blocks with asset

configuration c, $B = \{b_i^{c_i}, \ldots, b_n^{c_n}\}$, and different event types defined as sets containing at least one member for character animation events, $E_a = \{e_{a1}, \ldots, e_{am}\}$, visual effects, $E_v = \{e_{v1}, \ldots, e_{vk}\}$, and sound effects, $E_s = \{e_{s1}, \ldots, e_{sj}\}$, such that $E_a \cap E_v \cap E_s = \emptyset$ and $E = E_a \cup E_v \cup E_s$, the set $G = \{g_1, \ldots, g_n\}$, holds the occurrence frequencies for any event $e_i \in E$. Each occurrence frequency is given by the probability value $\{g_i \mid g_i \in \mathbb{Q}^+ \wedge g_i \leq 1\}$, that the event e_i will appear on any level block b, such that $\forall_i\, e_i \in E \wedge g_i \in G : \exists\, (e_i, g_i)$.

Since high probability values for any individual parameter could cause an explosion of content in the game, environment and creature events were capped to an occurrence once every 10 to 15 generation cycles (i.e., approximately once every 30-45 seconds).

At the beginning of the game session, each of the controllable game elements is initialized using a set of predefined values. However, these can be directly altered by the AID at any time according to the game state, or, in the biofeedback conditions, the player's emotional state (see Section 4:1). Thus, while the procedural content generator gives us real-time control over almost every aspect of the gameplay experience, the game director enforces the defined gameplay adaptations that drive the gameplay experience according to the game designer's vision.

Game Architecture

Vanish's architecture as a native, tailored implementation of the Emotion Engine

Although Vanish was fully developed in Unity, the game engine allows the creation of additional functionality via external application communication using scripts written in C#. This was a design requirement that we needed to incorporate into the game engine during its development to enable physiological interaction in Vanish. As such, we equipped the engine with our emotional recognition module – PIERS. In order to integrate it with Unity, we had to rewrite PIERS in C# setting it to retrieve the physiological data in real-time, from the NeXus-10 physiological data capturing hardware via a pooling mechanism. The game engine was then adapted to retrieve the computed AV ratings from PIERS (see Figure 4:5). In the biofeedback-adapted version of the game architecture, the NeXus-10 device is responsible for capturing physiological data for the SC, BVP and EMG channels. The Biotrace+ software suite then pre-processes the acquired data, derives the HR signal from BVP readings, and also performs the signal acquisition from the NeXus-10 device via Bluetooth.

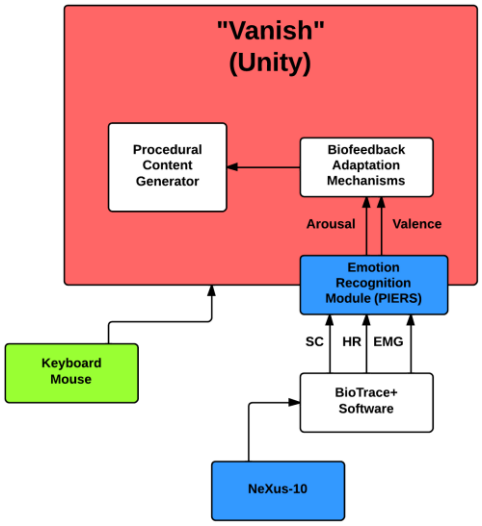


Figure 4:5. Biofeedback-adapted version of the game architecture. Notice that in this adapted version of the game's architecture, the AI director module of the control version is replaced by the biofeedback adaptation mechanisms, which determine the gameplay parameters.

PIERS was also responsible for, in parallel, converting the physiological data received from the BioTrace+ software suite into arousal and valence (AV) ratings, and feeding them to the biofeedback adaptation mechanisms module (which acted as a replacement for the AI director on the biofeedback versions of the game).

4.5 RESULTS

In this section, we present the statistical analysis performed on the collected player experience questionnaire answers and processed physiological measures (i.e., arousal/valence state estimates), segmented by type (player experience, physiological differences and condition preferences).

Our results provide evidence supporting statistical differences between each of the gaming conditions for several of the measured player experience dimensions (immersion, tension, positive affect, negative affect, and fun). They also show that, while the mean physiologically measured arousal and valence ratings did not vary significantly, a linear trend similar to the one observed in the subjective ratings is present. In light of the previous results, a detailed sub-group analysis was also performed, revealing significant differences between gender, player proficiency, and game genre preferences.

We conclude by presenting players' ability to identify each IBF condition (so as to assess the actual perceptible changes to the gameplay experience), their reported preferences regarding the three gaming conditions, and their subjective commentaries on the gameplay experience as a whole.

Player Experience Ratings: Statistical Analysis

Assessing interplay and dependencies between gameplay experience dimensions

To check whether there were any statistically significant inter-dependencies between the reported player experience dimensions, we performed a simple (Pearson) correlation analysis on this data. No differentiation was done between gaming conditions. Statistically significant correlations were observed for all gameplay dimensions (see Table 4:4), albeit with significant (up to ±100%) changes in their strengths.

Exploring these correlations grounded on the existing literature

Most of these correlations create an illustrative map describing how players perceive their gameplay experience, and are easy to explain. That immersion would correlate with flow and fun was to be expected, because it is generally indicative of a good gameplay experience (Ermi & Mäyrä, 2005). Immersion's positive correlation with competence and positive affect is also intuitive, because players that feel competent enjoy positive feedback from overcoming the game's difficulties; this is generally referred to as "Fiero" (Lazzaro, 2005). In turn, this positive feedback eases their absorption process into the game world, generating more immersion. The correlation between immersion and tension is also easy to explain, given that, as previously mentioned, horror games derive their immersive properties from the (tense) atmosphere, which is used to draw players into the game. Similarly, tension being heavily correlated with challenge indicates that players felt tenser as they felt more challenged and vice-versa. The weak correlation between Tension and competence has a slightly less intuitive explanation, but could hint that as more tension was put on the player, they felt more competent when able to properly respond to the obstacles. This seems to be partially corroborated by the mild correlation between competence and challenge, as well as flow, positive affect and overall fun. Regarding challenge, if we take into consideration Csíkszentmihályi's definition of flow as *"a state of concentration or complete absorption with the activity at hand and the situation where provided challenges and skill are perfectly balanced,"* and flow's known correlation with fun (Csíkszentmihályi, 2008), its correlations with flow and fun seem only natural. The same holds true for flow's (positive) correlations with positive affect and fun.

Table 4:4. Correlations (*r*-values, *p*-values) between reported Immersion (I), Tension (T), Competence (C), Challenge (Ch), Flow (Fl), Positive Affect (PA), Negative Affect (NA), and Fun (F) [13].

	I	T	C	Ch	Fl	PA	NA
T	**0.370 (0.002)**						
C	**0.318 (0.008)**	**0.249 (0.039)**					
Ch	0.224 (0.064)	**0.410 ($5*10^{-4}$)**	**0.322 (0.007)**				
Fl	**0.481 (0)**	0.187 (0.123)	**0.503 (0)**	**0.292 (0.015)**			
PA	**0.394 ($8*10^{-4}$)**	-0.133 (0.277)	**0.247 (0.041)**	0.062 (0.615)	**0.448 ($1*10^{-4}$)**		
NA	**-0.351 (0.003)**	0.167 (0.17)	-0.105 (0.389)	0.070 (0.567)	-0.229 (0.059)	**-0.694 (0)**	
F	**0.504 (0)**	**0.253 (0.036)**	**0.335 (0.005)**	**0.249 (0.039)**	**0.467 (0)**	**0.475 (0)**	-0.213 (0.079)

Gameplay experience dimension correlations as an explaining factor in game's emotional spectrum

These relationships appear to have non-orthogonal dependencies in several cases, which would indicate that some configurations of player experience (e.g., improve fun, while decreasing challenge for more novice players) might be difficult to achieve; thus, generating game design complications and/or limitations. However, given the nature of intrinsic motivation found in most games (overcoming a set of obstacles to reap some reward) these interrelationships may well be an inherent property of player experience. In some cases, they may not even make sense from a game design perspective (e.g., decreasing tension in a horror game, while increasing immersion and positive affect). In part, this is also constrained by the game's own limited emotional spectrum, which we present further into this section.

Order Effects on Physiological Input

While participants were given a 5-minute interval to rest between gaming conditions, which were randomized to avoid order effects on physiological – and thus emotional – readings, we wanted to confirm whether this protocol had the intended effect. As such, two repeated-

[13] For positive affect and negative affect, correlation pairs not achieving statistical significance ($p<0.05$) were tension, challenge, competence (NA only) and flow (NA only). Additionally the flow-tension ($p=0.12$) and challenge-immersion ($p=0.06$) pairs were also almost significant. All remaining correlations pairs were statistically significant ($p<0.05$). Critical *r-values* appear in bold.

measures analyses of variance (ANOVA) were conducted, using the different initial arousal and valence ratings for the first second of each game condition as the within-subject factor.

Assessing whether the experimental protocol was successful in mitigating order effects and experimental noise

For arousal ($\chi^2(2)=0.9553$, $p>0.05$) measurements, Mauchly's test showed that the sphericity assumption was met. However, for valence ($\chi^2(2)=0.6891$, $p<0.05$) it was violated. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity, $\varepsilon=0.7628$.

Statistical significance was not achieved for both the arousal ($F(2,44)=1.2747$, $p>0.05$) and valence ($F(1.52,33.56)=0.0827$, $p>0.05$) components, indicating that participants' emotional states were well matched at the start of all three gaming conditions, and thus no ordering effects were present in the physiological data.

Player Experience MANOVA Analyses

To better understand the impact of the gaming conditions on player experience, we conducted a MANOVA analysis, using the different game conditions as the within-subject factor and player demographics as between-subject factors. The metrics under evaluation were: the GEQ dimensions, overall fun and the average computed arousal and valence ratings over each gameplay session.

Did the biofeedback game conditions significantly impact players' perceived gameplay experience?

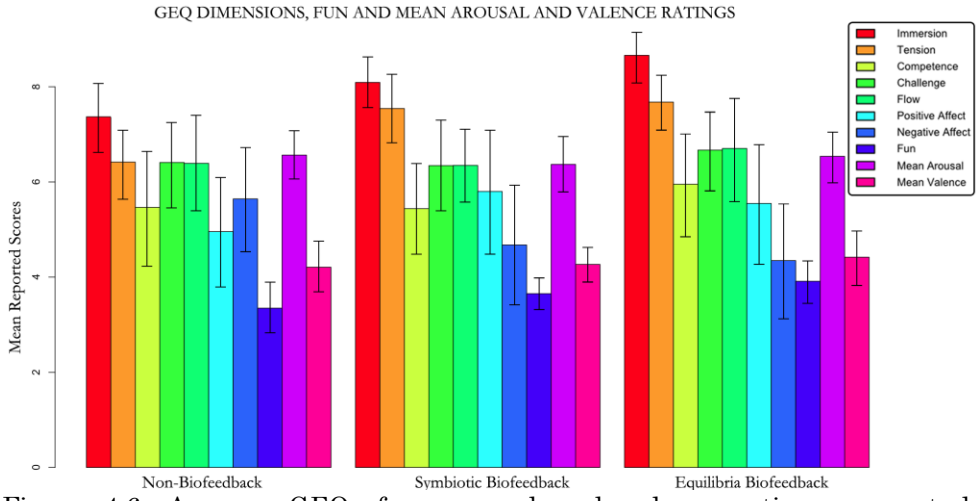


Figure 4:6. Average GEQ, fun, arousal and valence ratings over study participants for each gaming condition.

For GEQ components immersion ($\chi^2(2)=3.963$, $p>.05$), tension ($\chi^2(2)=0.356$, $p>.05$), competence ($\chi^2(2)=0.888$, $p>.05$), challenge ($\chi^2(2)=0.515$, $p>.05$), flow ($\chi^2(2)=1.775$, $p>.05$), positive affect ($\chi^2(2)=4.335$, $p>.05$), negative affect ($\chi^2(2)=2.639$, $p>.05$), as well as fun ($\chi^2(2)=1.954$, $p>.05$), and mean arousal ($\chi^2(2)=1.381$, $p>.05$), Mauchly's test showed that the assumption of sphericity had been met. For mean

valence ($\chi^2(2)=8.564$, $p<.05$) it was violated. Therefore, degrees of freedom were corrected using a Greenhouse-Geisser estimate of sphericity ($\varepsilon = 0.697$).

Exploring between-subjects effects and interactions for individual factors

No significant between-subjects effects were observed for all factors. There were, however, several statistically significant interactions for individual factors. Gamer type displayed an interaction with mean valence, where hardcore players (M=3.99, SD=1.15) reported slightly higher mean valence levels than casual players (M=3.63, SD=0.59); this could suggest that casual players are more easily unsettled by the negative aspects of horror games. This seems in line with the observed negative correlation between negative affect and immersion, further strengthening the idea that immersion is not necessarily elicited by a negatively affective experience. Along the same lines, players that liked horror games also reported considerably higher fun levels (M=3.95, SD=0.78) than players who did not (M=3.23, SD=0.95).

Additionally, two higher-order interactions were also found. Gamer type and genre preference affected reported negative affect for hardcore players that liked horror games; presenting the lowest negative affect scores (M=3.08, SD=2.62). Interestingly, they were followed by casual players that did not like horror games (M=5.09, SD=1.54) and casual players that did like horror games (M=5.93, SD=1.98). Finally, the demographic most affected by the game's negative theme were hardcore players that did not enjoy horror games (M=6.9, SD=0.69). These results would seem to indicate that while a positive or negative opinion of horror games influences players' susceptibility to negative emotions, player proficiency acts as a polarising/multiplicative factor. A similar interaction was detected in the gamer type and player gender effects for mean valence, with hardcore female players reporting the lowest valence (M=2.39, SD=0.52). This means they had the most negative experiences, followed by hardcore male players (M=4.14, SD=1.08), casual male players (M=4.6, SD=0.56) and casual female players (M=4.65, SD=0.61). Again, player proficiency seems to act as a divisive factor, especially for female players, who seemed remarkably susceptible to negative emotional states.

Table 4:5. Between-subjects effects summary.

Interacting Factors	*Measure*	*F-statistic*	*p-value*	*Partial* η^2
Gamer Type	Mean Valence	8.775	0.009	0.354
Genre Preference	Immersion	9.947	0.063	0.199
Genre Preference	Fun	5.721	0.034	0.253
Gamer Type * Genre Preference	Negative Affect	4.590	0.048	0.223
Gamer Type * Player Gender	Mean Valence	8.075	0.012	0.335

Regarding within-subjects effects, statistical significance (p<0.05) was found for immersion, tension, positive affect and negative affect, which indicates that the different mechanics on each game condition significantly affected the player experience on the most relevant factors for this game genre. The detailed statistics for these tests can be observed in Table 4:6. Additionally, there was a significant effect of gaming condition and game genre preference on the challenge factor (F=3.561, p=0.048, η^2=0.182), which hints that enjoying the game was a differentiating factor in at least one gaming condition.

Significant changes in Immersion, Tension and both affect dimensions across gameplay conditions were observed

Table 4:6. Player experience ratings statistical analysis. Left columns: *F*-statistic, *p*-value and partial *eta*-squared for reported game experience dimensions. Right columns: Statistical descriptors for each game experience dimensions over all gaming three conditions.

	MANOVA			Descriptive Statistics	
	F-statistic	*p-value*	*Partial* η^2	Mean	Standard Deviation
Immersion	8.214	**0.010**	0.339	8.04	1.28
Tension	3.871	**0.031**	0.195	7.21	1.46
Competence	0.473	0.627	0.029	5.62	2.22
Challenge	0.521	0.599	0.032	6.47	1.79
Flow	0.109	0.897	0.007	6.48	2.08
Positive Affect	3.647	**0.037**	0.186	5.43	2.63
Negative Affect	6.879	**0.003**	0.301	4.89	2.46
Fun	2.515	0.097	0.136	3.64	0.93
Mean Arousal	0.663	0.522	0.040	6.49	1.02
Mean Valence	0.504	0.545	0.031	4.30	0.98

Within-subjects Bonferroni contrasts were conducted to identify how the game conditions were differentiated. These revealed that for the component immersion, a significant (p<0.05) linear trend existed indicating that immersion increased linearly from the NBF (M=7.37, SD=1.48) to S-IBF (M=8.09, SD=0.99) and E-IBF (M=8.66, SD=1.07) conditions. However, it also revealed an even more significant (p<0.01) quadratic relationship between immersion and game conditions, which better explains the lower increase from S-IBF to E-IBF (0.57 points) than from NBF to S-IBF (0.72 points). Similarly — and in line with the identified correlation between immersion and tension — a quadratic trend was present in the tension component. Participants seemed to rate the E-IBF condition as the tenser of the three, with the S-IBF following closely in second and the NBF condition in last with a considerable difference of over 1 full point (see Table 4:7).

Post-hoc tests for statistically significant components show mostly quadratic trends, indicating that all conditions were different

Regarding the positive and negative affect components, quadratic trends were also observed. In terms of positive affect, the S-IBF condition reported — as expected — the highest positive affect and was closely followed by the E-IBF condition. The NBF condition ranked last. Surprisingly, despite being tailored to exacerbate players' emotional

states, it appears that the S-IBF condition actually had the opposite effect as players reported much lower negative affect ratings for it than for the NBF condition, which presents the most negative emotional experience. The E-IBF condition seems like it was able to effectively calm players (perhaps due to its player-aiding nature) and presented the least negative experience.

On the other hand, demographic and condition factors a linear trend was observed

Finally, regarding the effect of gaming condition and genre preference on the challenge factor, a linear trend was identified, which suggests that challenge increased linearly from the control condition to the biofeedback conditions for players that liked horror games and decreased, also linearly from the control condition to the biofeedback conditions for players that did not like horror games. This could be because players familiar with the genre were accustomed to typical mechanics, and experienced an adaptation period with the "new" mechanics in Vanish, thus causing an increase in challenge. Conversely, players not familiarised with the genre experienced simultaneous adaptation with every condition and were thus able to innately capitalise on the advantages of the biofeedback conditions. For convenience, the descriptive statistics and contrast results are summarised in Tables 4:7 and 4:8.

Table 4:7. Statistical data (mean, standard deviation) on reported gaming conditions/game experience dimension for significant within-subjects Bonferroni contrasts.

Measure	Demographic Factor	Game Condition		
		NBF (Control)	S-IBF	E-IBF
Immersion	None	(7.37, 1.48)	(8.09, 0.99)	(8.66, 1.07)
Tension	None	(6.42, 1.44)	(7.54, 1.20)	(7.68, 1.44)
Positive Affect	None	(4.96, 2.45)	(5.80, 2.70)	(5.55, 2.71)
Negative Affect	None	(5.64, 2.42)	(4.68, 2.42)	(4.35, 2.50)
Challenge	Positive Genre Preference	(5.72, 1.84)	(6.52, 2.03)	(6.84, 1.87)
	Negative Genre Preference	(7.16, 0.95)	(6.27, 1.48)	(6.45, 1.82)

Table 4:8. Within-subjects Bonferroni contrasts results for significant game experience measures.

Factors	Measure	Trend	*F*-Statistic	*p*-value	*Partial* η^2
Game Condition	Immersion	Linear	7.271	0.016	0.312
		Quadratic	8.679	0.009	0.352
Game Condition	Tension	Quadratic	7.047	0.017	0.306
Game Condition	Positive Affect	Quadratic	4.570	0.048	0.222
Game Condition	Negative Affect	Quadratic	10.395	0.005	0.394
Game Condition * Genre Preference	Challenge	Linear	6.325	0.023	0.283

Gaming Condition Impact on Physiological Metrics

Regarding the biofeedback mechanics' impact on players' emotional states, an empirical analysis of the relative density of players' emotional states over time (see Figure 4:7) reveals very dissimilar trends for the three game conditions. Overall, it seems that the implicit biofeedback condition (E-IBF) was able to more successfully balance players' emotional states by concentrating them on the central region of the AV space, while also reducing the density of emotional states on the AV space's 3rd quadrant (low valence and low arousal — i.e., reducing boredom). This seems in line with participant comments (see the participant opinion paragraphs further ahead on this sub-section) on the condition's more hectic and well-paced nature. Similarly, the explicit biofeedback condition was also able to reduce the density hotspot on the 3rd quadrant of the AV space, while also lowering player arousal and marginally increasing overall valence. This indicates that some players were indeed able to relax (to a certain degree) in the symbiotic condition (S-IBF), thus being able to achieve a self-perpetuating competitive advantage. In other words, relaxing improved their in-game abilities, which, in turn, further helped them to relax, creating a positive feedback loop. Some players were caught in the opposite direction (amplifying their own disadvantages by not relaxing), which justifies the incomplete reduction of some considerable hotspots on the 2nd and 3rd quadrants of the AV space.

Biofeedback conditions' impact on players' physiologically-measured emotional states

In conclusion, the diverging AV distributions presented by both of the biofeedback conditions (with respect to each other as well as the control condition) provide convincing evidence that biofeedback-augmented gameplay is suitable as both a dramatic enhancer and a regulator of player experience. Over the following sub-sections, we will explore in further detail the differences between specific demographic sub-groups. Additionally, a full break-down of players' emotional states over demographic and gaming conditions is presented in the appendix section of this thesis. Overall, the same patterns were observed in each sub-group; the Symbiotic (explicit) biofeedback condition widened the experienced emotional spectra, while also polarizing some extreme AV states (due to the aforementioned positive biofeedback loop), and the Equilibria (implicit) biofeedback condition balanced the emotional spectra even further, with no noticeable hotspots. The exception to this rule was female players for whom the implicit biofeedback condition seemed only to provide steering towards a neutral emotional state. Since we found no significant interactions between demographic factors, this could perhaps hint that female players react differently to implicit mechanics and thus require different slightly different guidelines when compared to male players.

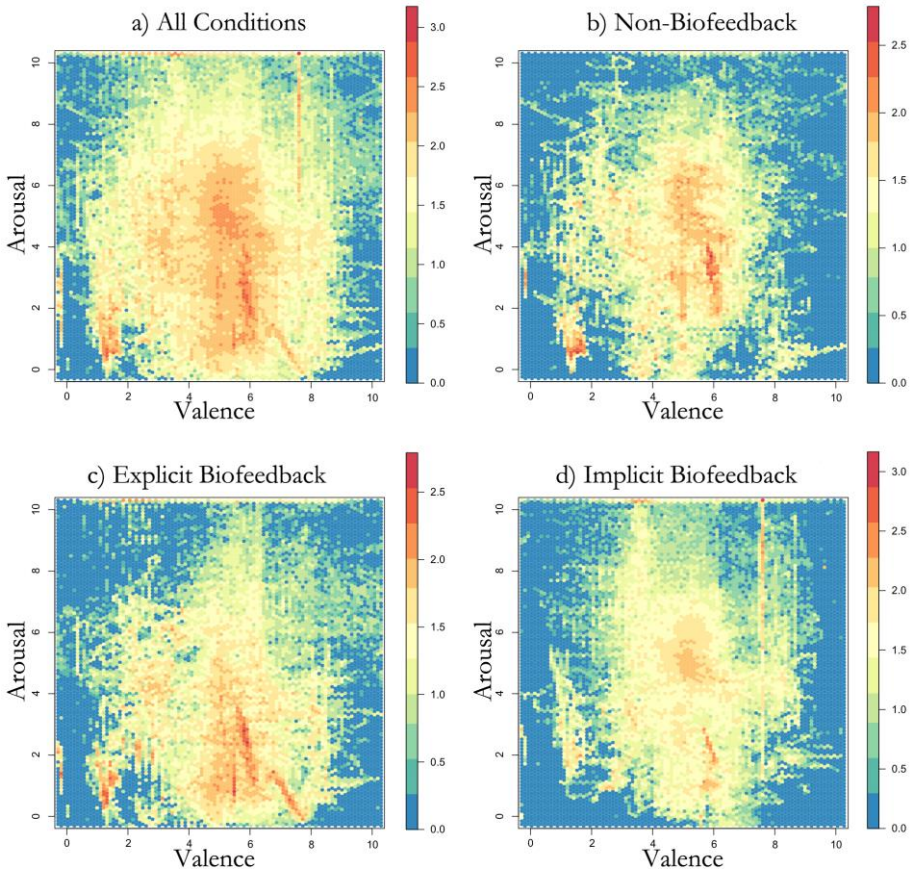


Figure 4:7. Normalised AV heat maps per gaming condition on a logarithmic scale over time (at a 32Hz sampling rate) spent on each emotional state.

Individual Player Spectra

Players' emotional spectra were very dissimilar but shared some fundamental group traits

Given that individual players presented noticeable differences in the collected metrics, we analysed whether this was also true for their respective individual emotional spectra. In fact, we found that virtually no two players' emotional spectra were alike (most were, in fact, quite dissimilar and centred on different areas of the AV space). The exception to this rule was that some spectra exhibited an almost two-dimensional binomial distribution; while others presented either a wide or very concise emotional range (these two properties were non-exclusive). An illustrative example of four participants' emotional spectra is shown in Figure 4:8.

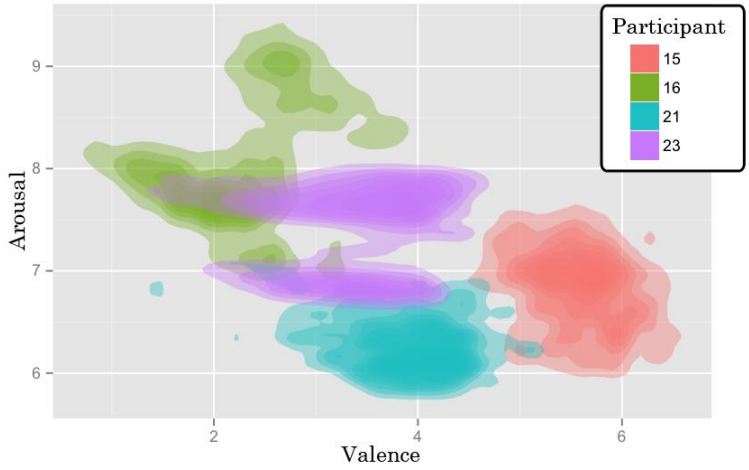


Figure 4:8. Illustrative AV density plots over all gaming conditions for four participants.

Group Spectra

Since our group analyses on player type revealed that casual players experienced significantly lower arousal and higher valence, we also analysed their combined emotional spectra (Figure 4:9). We wanted to understand if this was sufficiently perceptible, and if so, whether any other undetected patterns would emerge from the data. To our surprise, not only is this clearly evident even to the untrained eye, but the spectra are almost perfectly complementary. Independent two-tailed *t*-tests corroborate this – for average arousal ($t(41.42)=-4.04$, $p<.05$) and average valence ($t(25.89)=2.26$, $p<.05$). This seems to validate the common perception that hardcore players are less susceptible to emotional variations as a response to game stimuli (mean arousal (M=6.32, SD=0.82) and mean valence (M=4.57, SD=1.28)), than casual players (mean arousal (M=7.24, SD=0.77) and mean valence (M=3.90, SD=0.59)). In our opinion, this heavily suggests that, while different player types may report similar player experience metrics, the way they experience them is intrinsically distinctive and thus, both game adaptations (physiological or otherwise) and game design in general must take these factors into account, because players with different types may internally interpret the same emotional states differently.

These group traits were extremely evident between players with different proficiency levels

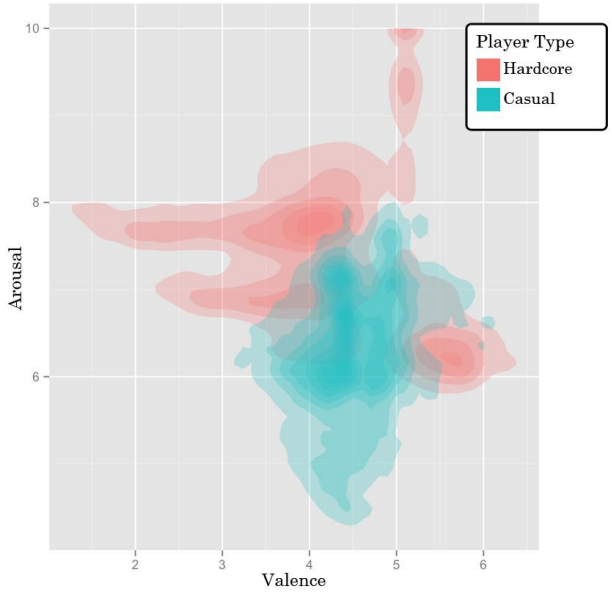


Figure 4:9. AV density plots per gamer type over all gaming conditions.

Segmenting players based on game genre preference also yields a predictable outcome

Segmenting players based on their game genre preferences yields a somewhat similar distribution (Figure 4:10), with players that reported enjoying horror games presenting an apparently more contained emotional spectrum – perhaps because of a habituation effect. Regarding player gender, we found no significant differences in the emotional spectra (Figure 4:10). Overall, female players seemed to exhibit a wider and more heterogeneous emotional spectrum, featuring more hotspots than their male counterparts, but no other significant differences are observable. Two isolated players, not constituting a characteristic feature of either gender population, generated the two hotspots in the second and third AV space quadrants. In line with these observations, no statistically significant differences were observed in these two groups.

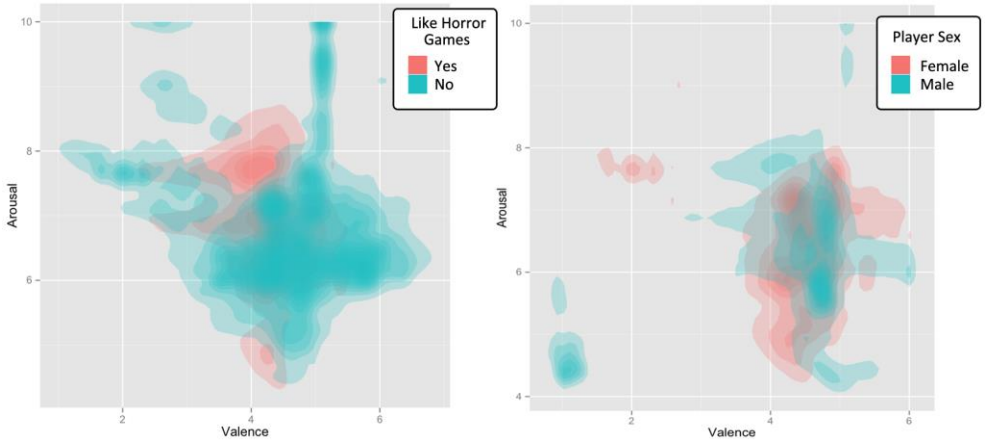


Figure 4:10. AV density plots segmented by players' game genre preference (left) and gender (right).

Condition Preference and Identification

As we have previously mentioned, players were asked to play each gaming condition in a Latin Square order to avoid order effects. Thus, we were also interested in: *a)* whether players were able to notice differences between the gaming conditions, and *b)* which conditions players preferred.

When asked about this, all but one player said that they noticed that there were differences between the conditions. While some players reported that the game pacing and character's behaviour seemed to respond to their emotional state, they were unable to accurately pinpoint how the game changed (i.e., the specific biofeedback gameplay mechanics). Thus, the 23 players that reported noticeable differences in gaming conditions were briefed on each condition's gameplay mechanics and asked to identify them according to their particular gaming order.

Despite condition order being blind, most players were able to correctly identify them

As can be seen in the confusion matrix, almost half (12) successfully identified every condition. Only two players erred on the recognition of all conditions. The remaining 9 participants correctly identified one condition but mismatched the remaining two; mostly the control and S-IBF conditions (6 out of 9).

The reason behind this might be that some of the S-IBF mechanics are difficult to compare, because the player does not have a base value with which to compare (i.e., for how long is the character usually able to run?). It also suggests that not all players are fully aware of their own emotional state while playing, which prevents them from correctly assessing the mechanism. S-IBF mechanics are supposed to be transparent to the player. Therefore, a simple solution for this confusion in commercial titles would be to brief players on the mechanic — something that, because of our research questions, was not advisable in this case.

Gaming Conditions	NBF (Predicted)	S-IBF (Predicted)	E-IBF (Predicted)
NBF (Actual)	12	7	3
S-IBF (Actual)	7	14	2
E-IBF (Actual)	3	2	17

Confusion matrix for gaming condition identification. The matrix is symmetrical because the two participants that misidentified all gaming conditions balanced each other's false positive and negative counts.

Given the high number of correct assessments by players, it would thus seem that despite some players' lack of awareness of their own emotional states, our design guidelines for both biofeedback conditions were appropriate, resulting in concrete, noticeable differences and distinctive player preferences between conditions.

Condition Preferences

Players showed a clear preference for the biofeedback conditions

Additionally, upon attempting to identify each gaming condition, players were asked to order them by preference. Players answered these questions by themselves and were not given any feedback on whether they correctly identified the conditions. This was done to avoid any biasing effects. Overall, participants' opinions strongly leaned towards E-IBF, which registered 43% of their preferences. The S-IBF condition came in second place, with 39% of participants' preferences. The control condition was the least preferred one, with only 13% of the votes (see Figure 4:11). We find these to be very positive results, since they clearly suggest that the biofeedback conditions presented some form of added value towards the gaming experience. Participants' slight inclination towards E-IBF when compared with S-IBF might suggest that S-IBF is difficult to control and players prefer to relinquish their control to the E-IBF mechanisms instead.

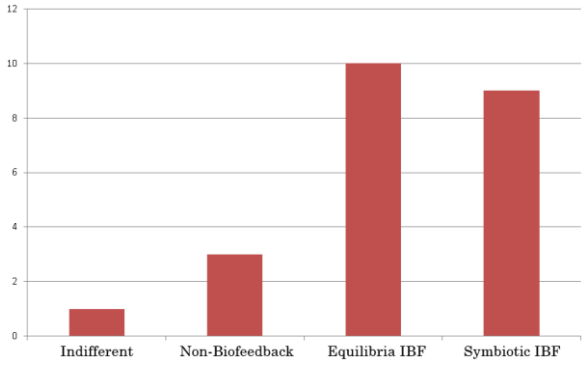


Figure 4:11. Participant preference per gameplay condition.

Participant Opinions

When asked about the potential of biofeedback-augmented games (not limited to horror games), participants agreed unanimously that it positively contributed towards the gaming experience.

Regarding the E-IBF condition, two participants felt like "*the game was being condescending"* or *"toying"* with them (P3). For participant 17,

this was a particularly amusing by-product of the game's procedural generation[14]:

"I found a dead end, and when I turned back and took the same path, it was all different, for moments I thought it was me, but no, the map really changed" (P17).

Additionally, regarding the game's procedural generation mechanics, two participants (P13, P21) suggested that the evasion tunnel's purpose should either be explained or be more self-evident:

"I saw the little tunnels but I never had the curiosity of going in there; it was pretty dark and I did not know what was inside" (P13).

After being briefed on its purpose and generation procedure participant 21 mentioned that:

"When I was running from the creature all I could do was look ahead and press the sprint key, if I knew I could hide from it on those tunnels I would probably search for them" (P21).

Players appeared to especially enjoy the biofeedback conditions' "mind reading" abilities and more dynamic pacing

Some players also expressed some frustration towards the creature AI, claiming it was somewhat *"unfair:"*

"When the creature started appearing, it was always running from me, so I kept running into it (thinking it wasn't a threat), when suddenly it lunged at me! I found it really unfair since there was no previous notice of it" (P19).

Nevertheless, most participants confirmed that the E-IBF version was indeed the one where they felt more *"immersed"* (P6) in the game and felt like it was *"driven by a purpose"* (P14), because of the timing chosen for the events to occur:

"It totally caught me off guard, I was very agitated on the beginning of the game, but the game pacing became progressively slower, and when I felt more relaxed, the creature appeared and started chasing me!" (P14).

Regarding the S-IBF condition, several participants stated that they felt like the game mechanics were intensifying their emotional state:

[14] This was an intended feature, as dynamic level layouts are somewhat common features in survival horror games. Since the game revolves heavily around the game character's sanity (or lack thereof), we decided to correlate this event with the player character's mental deterioration, thus helping to convey this notion to the player and, at the same time, tap into his own psyche in an attempt to subconsciously link the two.

"When the character started gasping and I could hear his heart beating faster and louder, it made me even more agitated" (P10).

Similarly to the other biofeedback condition, there were also some gameplay mechanism improvement suggestions, mainly for the anaerobic metabolism attributes and sanity mechanics. Participant 6 had the following comment — and solution proposal — on the anaerobic attributes mechanic:

"I could not perceive if there was a difference on the running speed because I was always running and the character was continuously exhausted" (P6).

"It would be easier if there was a stamina gauge on the user interface, so the player can have a better perception of how much stamina the character has left." (P6).

Concerning the sanity mechanic, participant 5 triggered the fainting mechanism and had the following comment/solution on its perceptibility:

"When the character fainted, for moments I thought I had lost the game, because I could not see or hear anything" (P5).

"A ringing sound when the screen is black would probably hint the player of what is happening with the character" (P5).

4.6 DISCUSSION

In this chapter we presented a study on the effects of what we consider explicit (Symbiotic) and implicit (Equilibria) indirect biofeedback techniques on games. Based on the lessons learned from this research, our main findings — with respect to our original research questions — are the following:

Static indirect biofeedback is capable of modulating players' emotional states

- Q1: *Can we modulate the "scary" experience of playing a horror game using physiological sensors and real-time processing*? Although players experienced distinct emotional spectra while playing the game (see Figures 4:8 and 4:10), both biofeedback conditions were able to successively shift them according to the described design guidelines. This resulted in each biofeedback condition eliciting considerably different global emotional state distributions over all participants (see Figure 4:7). In our opinion, both the wide range of observed emotional spectra and the biofeedback mechanics' ability to shift them independently present convincing evidence that biofeedback-augmented

gameplay is not only able to modulate players' affective experience in horror games, but that it is also suitable for emotionally-regulated games aiming to elicit much more complex affective experiences (e.g., targeting different emotional states over time or in response to specific events/game locations).

- *Q2: Do biofeedback-enabled adaptive mechanics have a significant impact on the players' gameplay experience?* Both biofeedback conditions offered players a more immersive, tense, and emotionally-rewarding experience, all of which are key dimensions of the gaming experience in horror games. This was evident in players' opinions and (blind[15]) preferences of each gaming condition. Biofeedback control was almost unanimously (only one participant disagreed) perceived as adding extra value to the game by increasing the gameplay depth. Furthermore, all participants recognised its potential in future applications, both in similar and different game genres.

It also delivers tangible value in terms of subjective user experience

- *Q3: How do different types of indirect biofeedback mechanics compare to each other in terms of user preferences and experience?* Participants preferred the Equilibria IBF version of the game to the other biofeedback version (S-IBF). Their comments on both versions suggest that the Symbiotic IBF version was difficult to control in such a stress-laden atmosphere and that the gameplay mechanics were not always obvious (perhaps an in-game sanity meter in future studies where condition blindness is not required would help alleviate this issue). On the other hand, the Equilibria IBF version was considered slightly more appealing, mainly because some players perceived its level generation and progression mechanisms as more advantageous.

It would seem both biofeedback variants have their advantages, with players still preferring less control over indirect mechanics

- *Q4: Do different types of players (distinguished by sex, proficiency, and genre preference) present any noticeable differences in how they experience these modifications?* Different types of players (based on their sex, proficiency, and genre preferences) interpret several aspects of the gameplay experience in significantly different ways (see Section 4:3 and 4:4 for a thorough discussion on these effects). This gives credibility to the notion that specific design guidelines could be developed according to player types.

Player proficiency should be the main factor in adapting different game design guidelines or gameplay adaptations

[15] As we previously mentioned, players were fitted with the physiological apparatus on all gaming conditions to avoid bias effects.

Additionally, we found that:

Players are not aware of their emotional states and games are bound by their intrinsic emotional spectra

- Players dissociate from their own emotional states while playing. However, this dissociation is not easily measurable and, therefore, it is difficult to map a player's emotional state to in-game representations.
- There are various orthogonal relationships between player experience dimensions, some more evident than others (e.g., immersion and tension, as opposed to tension and flow). These would indicate that not all configurations of player experience might be attainable, while others may require more complex game design guidelines.

Throughout the following section, we will discuss the impact of both biofeedback conditions on player experience, the significance of the observed emotional spectra, and the relative advantages of both biofeedback conditions. We then conclude with possible improvements, limitations and an analysis of our renewed research focus.

Impact and Effectiveness of Affective Biofeedback Gameplay Mechanics

The immersion, tension and positive/negative affect metrics present strong evidence that players enjoyed playing the game more using either form of biofeedback than without biofeedback. This was especially evident in players, who had previous experience with the genre and enjoyed survival horror games. These attributes appeared to further intensify the experience. In our opinion, there are at least three factors that may contribute towards this outcome.

Novelty is an ever-present factor in biofeedback research which can only be addressed in time with larger-scale studies

The first of these is the technology's novelty factor: Only a couple of participants had previously heard about biofeedback and while they understood the concept, none of them were aware of the different types or how they worked. Participants were given no information about the biofeedback capability of each condition. However, they were evidently able to discern differences between the conditions, because most players correctly identified the gaming conditions. Previous studies (Dekker & Champion, 2007; Kuikkaniemi et al., 2010; L. E. Nacke et al., 2011) have discussed that it is possible that this can imprint a bias (inflation) effect on players' ratings. The only way to assess whether this could affect the observed metrics would be to carry out a similar study over an extended period of time, preferably involving several games to mitigate game design issues. Unfortunately, the potential to conduct such research is limited both by the small collection of currently available

biofeedback games and the cost of physiological sensors. Properly (re)-calibrating the sensors over such a long time period and having enough devices to run parallel sessions would also pose some logistical issues, which would further complicate the study design. We discuss these issues further into this section.

The second factor that we feel might impact players' ratings is related to the two immersion theories described by (Brown & Cairns, 2004) and (Ermi & Mäyrä, 2005), respectively. Brown et al.'s definition of "Engrossment" states that it is achieved "*when game features combine in such a way that the gamers' emotions are directly affected by the game*;" Ermi et al.'s challenge-based immersion is "*the feeling of immersion when one is able to achieve a satisfying balance of challenges and abilities*" and "*can be related to motor or mental skills*" (Ermi & Mäyrä, 2005). Since, in both biofeedback conditions, players' emotions were used to drive the game's progression and to challenge their mental relaxation (S-IBF), as well as match their gaming abilities (E-IBF), it would seem natural that immersion (along with the other components that comprise a horror game, such as tension and player affect) would be affected. While most participants commented that the biofeedback mechanics added depth to the gameplay experience, making it feel more intense, they also mentioned that, in retrospect it seemed somewhat eerie. In this particular genre, this served to contribute beneficially to the game's atmosphere, because it made them face the game as an intelligent entity, rather than simply a random event generator.

Leveraging players' emotional states to drive game progression created a more complex atmosphere and thus stronger bond with the game world

The third factor that we felt might explain some of the observed differences between biofeedback conditions is that each of them altered dissimilar gameplay aspects (i.e., level generation vs character attributes). Despite our efforts to balance the number of occurring gameplay adaptations during play sessions, it is possible that these two gameplay aspects have intrinsically different effects on players' psyches and thus some variability on the observed effects can be attributed to the game design. In our opinion, this constitutes an important design guideline for future games and studies – quantify the effects of each gameplay mechanic in early versions of the game design for future reference during the game-balancing phases.

Symbiotic (Explicit) vs. Equilibria (Implicit) Indirect Biofeedback Game Design

Since both biofeedback conditions introduced a novel interaction paradigm that presented players with actionable advantages, it is not surprising that players preferred them to the control condition. Not surprisingly, the game's difficulty seemed to decrease from the control

condition to the Equilibria and Symbiotic IBF conditions (Figure 4:6), though this comparison did not achieve statistical significance.

Higher success rates and a sense of achievement benefited the E-IBF condition in the preference ratings

Based on these findings, we believe that the increased enjoyment reported by players in the Equilibria IBF condition was also affected by the feeling of accomplishment encouraged by this condition. An interesting conclusion was that, although the game decreased its difficulty for players who were struggling, this was interpreted as a reward for their efforts, hinting that players felt entitled to some form of achievement for their efforts. This goes hand in hand with the intrinsic philosophy behind video games (overcome an obstacle to receive a reward).

This feeling of accomplishment was also mentioned by some of the players who managed to calm down during the S-IBF gameplay session, but was not so pronounced due to the Symbiotic IBF mechanics' less discernible nature. In fact, we believe players who reported this increased sense of accomplishment were only able to do so because the positive feedback cycles elicited by this condition created clearly noticeable alterations to gameplay and physiological states.

Gray borders between IBF mechanic types and their exploitability

An implicit IBF mechanic may shift into an explicit IBF mechanic, if players become aware of it. While we feared that this would make the biofeedback mechanics susceptible to exploitation (and potentially increase the previously mentioned positive feedback cycles), from our analysis, a mixed approach seems to be a better choice, because a purely implicit IBF implementation does not readily offer enough feedback for players to take full advantage of its potential. On the other hand, the possibility of physiological hacks or cheats is a matter that must be researched in the long-term and presents various interesting questions and challenges for game designers targeting physiological interaction.

The biofeedback conditions affected player experience differently in terms of immersion and tension, because players rated these higher for the Equilibria IBF condition than for the Symbiotic IBF condition. This can perhaps be explained by the correlation factor (r=0.41) found between challenge and tension and the latter's subsequent correlation with immersion (r=0.37), which might suggest a bleeding effect that warrants future investigation.

Based on the obtained feedback and our own empirical analysis, we believe that similarly to how direct and indirect biofeedback should have different application focuses (L. E. Nacke et al., 2011), the same holds true for the Symbiotic and Equilibria indirect biofeedback subtypes. In our opinion, the two biofeedback types can be used in a complementary fashion; Symbiotic (explicit) indirect biofeedback can be

used to motivate players to play within a specific gameplay style, rewarding them for doing so, Equilibria biofeedback can be used in a more opaque manner to produce gameplay adaptations that enable players to make decisions regarding gameplay style, or adapt player experience in a way that makes those decisions more or less relevant, depending upon the logic of the game. Obviously, they can also be used independently, in this case, with the implicit mechanics serving as an emotional regulator/enforcer of the game designer's vision of the gameplay experience.

4.7 SUMMARY

Our study shows evidence in favour of augmenting game mechanics with affective physiological data. However, we focused on a specific game genre to perform a deeper analysis following previous investigations (Dekker & Champion, 2007; Kuikkaniemi et al., 2010; L. E. Nacke et al., 2011; Rani et al., 2005), and thus our work is limited in its applicability to other game genres.

Scope limitations include game genre and (despite mitigated) novelty effects

Additionally, although we explored various game mechanics and two different indirect biofeedback types, different choices with regard to game mechanics and biofeedback would most likely result in different player experience results. Furthermore, despite providing enough time for players to clearly notice differences in the gameplay experience, the playtime used was still not sufficient to account for novelty and habituation effects to this new technology. To provide a more complete investigation of this technology, these effects must be evaluated in future long-term studies.

From a perspective related to biofeedback-augmented games, this technology's most significant limitations are: *1)* the game designer's willingness to incorporate physiological interaction into their core gameplay mechanics, and *2)* the physiological sensors' proper calibration. Usually, the availability of the necessary hardware is mentioned as the primary challenge of using this technology, but we feel that this argument is losing momentum with the introduction of low-cost solutions like BITalino and the expression of interest in such technologies by industry giants such as Valve. Regarding sensor calibration, a potential solution would be to borrow the attributes of natural interaction solutions (e.g., Kinect and WiiMote) and gamify the calibration process, effectively masking it within the game's initial tutorial sections (Flatla, Gutwin, Nacke, Bateman, & Mandryk, 2011).

The main practical limitations are incorporating physiological interaction and sensor calibration

Fundamentally, the main limitation is the system's inability to learn from players' reactions

More broadly, the final (and perhaps most important), limitation intrinsic to static biofeedback techniques is their inability to learn from players' reactions. The aforementioned novelty and player habituation effects may pose a serious limitation concerning the long-term potential and relevance of biofeedback technology in a gaming context.

A potential solution to this issue would be to, given the distinctive traits presented by different types of players, shift between a set of predefined game adaptation mechanics as players become more accustomed to the game. These could, for example, be measured through in-game metrics or specific achievement types (e.g., once a player reaches a certain level of proficiency, measured by earning a certain achievement, the game's adaptation mechanics shift towards a new, more appropriate, predefined set).

However, this could be viewed as a work-around to the actual problem and thus falls short of our vision. Because we believe that biofeedback can be used to induce target emotional states, we are interested in exploring how it can be used to enforce a specific affective experience that embodies the game designer's original vision. In our opinion, doing so requires a detailed model of how each player reacts and the idiosyncrasies that make him unique: an affective reaction model.

A renewed research focus on creating and testing models of players' emotional reactions

In this chapter, we have shown that player experience can be influenced in a statistically significant way by the integration of basic (static) biofeedback mechanisms (thesis objective IV). Leveraging the collection of a wide body of data from our study, we also seized the chance to do this analysis in a more thorough, detailed manner spanning multiple points of view (objective vs. subjective); a study that from our analysis was missing in the literature. Additionally, while not a thesis objective *per se*, we presented and formalised our purpose-built case study.

Having collected players' emotional states and game event logs for all three gaming conditions, we now turn our attention towards modelling players' emotional reactions as closely as possible. Thus, Part VI will have a tripartite focus: *1)* identify and extract players' emotional reactions from the emotional state and game event logs, *2)* use the emotional reactions to model players' reactions – both individually and as a group, and *3)* exhaustively assess whether the constructed models can be used to shift players' emotional states in a predetermined manner using a symbolic simulation of Vanish's game world.

REFERENCES FOR CHAPTER IV

Ambinder, M. (2011). Biofeedback in Gameplay: How Valve Measures Physiology to Enhance Gaming Experience. In Game Developers Conference.

Bernhaupt, R., Boldt, A., & Mirlacher, T. (2007). Using emotion in games: emotional flowers. In Proceedings of the international conference on Advances in computer entertainment technology (ACE) (pp. 41–48). doi:10.1145/1255047.1255056

Bersak, D., McDarby, G., Augenblick, N., McDarby, P., McDonnell, D., McDonald, B., & Karkun, R. (2001). Intelligent biofeedback using an immersive competitive environment.

Blanchard, E. B., Eisele, G., Vollmer, A., Payne, A., Gordon, M., Cornish, P., & Gilmore, L. (1996). Controlled evaluation of thermal biofeedback in treatment of elevated blood pressure in unmedicated mild hypertension. Biofeedback and Self-Regulation, 21(2), 167–190.

Bourg, S. (2004). AI for Game Developers. O'Reilly & Associates.

Brown, E., & Cairns, P. (2004). A grounded investigation of game immersion. In Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04 (p. 1297). New York, New York, USA: ACM Press. doi:10.1145/985921.986048

Bryant, M. A. M. (1991). Biofeedback in the treatment of a selected dysphagic patient. Dysphagia, 6(2), 140–144.

Cavazza, M., Pizzi, D., Charles, F., Vogt, T., & André, E. (2009). Emotional input for character-based interactive storytelling. In Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1 (pp. 313–320). International Foundation for Autonomous Agents and Multiagent Systems.

Csíkszentmihályi, M. (2008). Flow: The Psychology of Optimal Experience (p. 336). HarperCollins.

Dekker, A., & Champion, E. (2007). Please biofeed the zombies: enhancing the gameplay and display of a horror game using biofeedback. In Situated Play, Proceedings of the Digital Games

Dong, Q., Li, Y., Hu, B., Liu, Q., Li, X., & Liu, L. (2010). A Solution on Ubiquitous EEG-based Biofeedback Music Therapy. In IEEE 5th International Conference on Pervasive Computing and Applications (ICPCA) (pp. 32–37). doi:http://dx.doi.org/10.1109/ICPCA.2010.5704071

Ermi, L., & Mäyrä, F. (2005). Fundamental components of the gameplay experience: Analysing immersion. In Digital Games Research Association Conference: Changing Views - Worlds in Play.

Flatla, D. R., Gutwin, C., Nacke, L. E., Bateman, S., & Mandryk, R. L. (2011). Calibration games: making calibration tasks enjoyable by adding motivating game elements. In Proceedings of the 24th annual ACM symposium on User interface software and technology (pp. 403–412). doi:10.1145/2047196.2047248

Gilleade, K. M., Dix, A., & Allanson, J. (2005). Affective Videogames and Modes of Affective Gaming : Assist Me , Challenge Me , Emote Me. In DIGRA - Digital Games Research Association (pp. 1–7).

Hjelm, S. I., & Browall, C. (2000). Brainball–Using brain activity for cool competition. In Proceedings of NordiCHI (pp. 177–188).

Huang, H., Ingalls, T., Olson, L., Ganley, K., Rikakis, T., & He, J. (2005). Interactive multimodal biofeedback for task-oriented neural rehabilitation. In 27th Annual International Conference of the Engineering in Medicine and Biology Society (IEEE-EMBS) (pp. 2547–2550).

Hudlicka, E. (2009). Affective Game Engines : Motivation and Requirements. In International Conference on Foundations of Digital Games. Orlando, Florida, USA.

Ijsseltein, W. A., Poels, K., & De Kort, Y. (2008). The game experience questionnaire: Development of a self-report measure to assess player experiences of digital games: FUGA technical report, Deliverable 3.3.

Kim, J., Bee, N., Wagner, J., & André, E. (2004). Emote to win: Affective interactions with a computer game agent. In GI Jahrestagung: (1) (pp. 159–164).

Kuikkaniemi, K., Laitinen, T., & Turpeinen, M. (2010). The influence of implicit and explicit biofeedback in first-person shooter games. In Proceedings of the 28th international conference on Human factors in computing systems (pp. 859–868).

Lazzaro, N. (2005). Why We Play Games : Four Keys to More Emotion Without Story. Design, 18, 1–8. doi:10.1111/j.1464-410X.2004.04896.x

Loveridge, S. (2014). Sony confirms biometric sensors were tested for the PS4 controller. TrustedReviews. Retrieved March 31, 2014, from http://www.trustedreviews.com/news/sony-confirms-biometric-sensors-were-tested-for-the-ps4-controller

Mandryk, R. L., Dielschneider, S., Kalyn, M. R., Bertram, C. P., Gaetz, M., Doucette, A., … Keiver, K. (2013). Games as neurofeedback training for children with FASD. In Proceedings of the 12th International Conference on Interaction Design and Children (IDC '13) (pp. 165–172). doi:10.1145/2485760.2485762

Marshall, J., Rowland, D., Egglestone, S. R., Benford, S., Walker, B., & McAuley, D. (2011). Breath control of amusement rides. In Proceedings of the SIGCHI ACM Conference on Human Factors in Computing Systems (pp. 73–82).

Nacke, L. E., Kalyn, M., Lough, C., & Mandryk, R. L. (2011). Biofeedback Game Design : Using Direct and Indirect Physiological Control to Enhance Game Interaction. In Proceedings of the 2011 annual conference on Human factors in computing systems (pp. 103–112). ACM.

Narcisse, E. (2013). Kinect 2.0 Sees Your Face, Muscles and Soul. Maybe Not That Last One. Kotaku.

Negini, F., Mandryk, R. L., & Stanley, K. (2014). Using Affective State to Adapt Characters, NPCs, and the Environment in a First-Person Shooter Game. In IEEE Games, Entertainment, and Media (p. Too appear). Toronto, Canada.

Parnandi, A., & Gutierrez-Osuna, R. (2014). A comparative study of game mechanics and control laws for an adaptive physiological game. Journal on Multimodal User Interfaces. doi:10.1007/s12193-014-0159-y

Parnandi, A., Son, Y., & Gutierrez-Osuna, R. (2013). A Control-Theoretic Approach to Adaptive Physiological Games. In Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on (pp. 7–12). IEEE. doi:10.1109/ACII.2013.8

Pigna, K. (2009). Sony Patents "Emotion Detecting" PS3 Technology. 1Up.com. Retrieved from http://www.1up.com/news/sony-patents-emotion-detecting-ps3

Pope, A. T., Stephens, C. L., & Gilleade, K. (2014). Biocybernetic Adaptation as Biofeedback Training Method. In Advances in Physiological Computing (pp. 91–115).

Rani, P., Sarkar, N., & Liu, C. (2005). Maintaining optimal challenge in computer games through real-time physiological feedback. In Proceedings of the 11th International Conference on Human Computer Interaction (pp. 184–192).

Reynolds, E. (2013). Nevermind. Retrieved from http://www.nevermindgame.com/Nevermind_Game/Nevermind__The_Game.html

Riva, G., Gaggioli, A., Pallavicini, F., Algeri, D., Gorini, A., & Repetto, C. (2010). Ubiquitous Health for the Treatment of Generalized Anxiety Disorders. In UbiComp '10. Copenhagen, Denmark.

Rocchi, L., Benocci, M., Farella, E., Benini, L., & Chiari, L. (2008). Validation of a wireless portable biofeedback system for balance control: preliminary results. In IEEE Second International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth (pp. 254–257).

Schaefer, A., Nils, F., Sanchez, X., & Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films : A new tool for emotion researchers. Cognition and Emotion, 24(7), 1153–1172.

Stepp, C. E., Britton, D., Chang, C., Merati, A. L., & Matsuoka, Y. (2011). Feasibility of game-based electromyographic biofeedback for dysphagia rehabilitation. In 5th International IEEE/EMBS Conference on Neural Engineering (NER) (pp. 233–236).

Chapter V

EMOTIONAL REACTION TRIANGULATION

TRIANGULATING PLAYERS' EMOTIONAL REACTIONS: A PSYCHOPHYSIOLOGICAL METHOD USING DIGITAL MEDIA STIMULI

OUTLINE

In this chapter we take the first step towards the creation of players' affective reaction models; defining a method for the extraction of emotional responses to media stimuli. We target this issue by presenting a method capable of automatically annotating and triangulating players' physiologically interpreted emotional reactions to in-game events.

Given that current affective user experience studies require laborious and time-consuming data analysis, as well as dedicated affective classification algorithms, we chose to take this chance to embed this method into a standalone, open-sourced tool. With this we hope to leverage/disseminate our previous efforts in developing an affective classification algorithm, streamline the annotation process and ultimately contribute towards the comparability of the obtained results. We also expect this tool to contribute in future studies by providing both a deeper and more objective analysis on the affective aspects of user experience.

Throughout this chapter we describe the development and benefits presented by our tool, which include: enabling researchers to conduct objective a posteriori analyses without disturbing the gameplay experience, automating the annotation and emotional response identification process, and formatted data exporting for further analysis in third-party statistical software applications.

Due to their high emotional elicitation potential, digital games have been increasingly used in many tangent research areas. For example, digital games are a suitable alternative to real life studies or dangerous, expensive or logistically challenging studies, such as phobia treatment (R. Mandryk & Atkins, 2007). Common methods used in these types of research mostly include behavioural observation, psychometric questionnaires or psychophysiological data annotation. Most studies still use the first two former methods, which offer rather vague insights into the emotional alterations that the presented stimuli elicited on each participant. More recent studies have recently started exploring the potential of psychophysiological data (RL Hazlett, 2006; Leon et al., 2007; R. Mandryk & Atkins, 2007; L. E. Nacke et al., 2010), but usually

require manual sensor calibration and annotation of each reaction. This means there is a current need for a method capable of automating the classification of this psychophysiological data. Also, due to the high workload involved in the required data annotation (e.g. logging game-related events and identifying relevant emotional responses), this process should also be automatized.

The need for a psychophysiological data analysis tool in affective human-computer studies

While developing a standalone solution for this issue was not *sine qua non* for attaining the objectives set forth for this thesis, it is a recurring issue in the literature. Effectively, a manual end-to-end data annotation and analysis process can take up tens or even hundreds of man-hours in a relatively small study. To make matters worse, any emotion recognition method used is likely to be a redesigned, re-implemented or new variant of existing methods, which further hinders any attempt at an objective comparative analysis of the presented results. Thus, in the spirit of "scientific cooperation, we chose to develop this tool and open-source it [16].

Our proposed solution consists of a tool to automate the analysis of psychophysiological-measured emotional reactions to in-game stimuli. This tool allows users to load gameplay session videos and synchronise them with the corresponding physiological recordings. Rather than displaying the raw physiological readings, as with more traditional approaches, the tool then interprets these readings as a continuous prediction of arousal and valance ratings (Russel, 1980b) using our previously proposed method (see Chapter III). This continuous prediction is built-in as a standalone module that can be replaced or parameterised on-the-fly by users so that the tool is useful in a wide range of application scenarios. The tool then allows users to replay the game session videos and manually annotate gameplay events using a simple GUI. Alternatively, users can also import text-formatted game event logs in case these are available, thus bypassing (or augmenting) the manual annotation phase altogether.

Defining a data processing workflow based on existing practises and protocols

Once users have finished the annotation of game events, the tool automatically estimates which of the annotated game events prompted emotional reactions. Given that this "emotion-event triangulation" process is the cornerstone of the tool, we chose to adopt it as its name (EET). The triangulation itself is done by automatically isolating the emotional responses to each annotated event by applying a parameterised two-dimensional local maxima/minima search algorithm to the emotional classification signal (i.e. the arousal/valence signal result from the predictive model). The system also allows subjective

[16] Contextualising the proposed work in terms of the E² architecture, the emotion extraction and identification module implements the same functionality of ARE²S, minus the model creation.

metrics, such as questionnaire responses or user commentaries to be added for each individual event. Sessions can also be saved to a serialized .eet data file that can be loaded in a posterior point in time, should a gameplay session require additional analysis. Finally, we included an export feature that writes the identified emotional responses to a structured (tab-delimited) text file for further analysis in third party software tools such as R, Weka or SPSS, as is usual in many of these studies.

After we have discussed all of the tool's features we also present a validation on its automatic emotional response detection capabilities. This is meant as a measure of the system's overall adequacy for our needs and how it contributes to the typical psychophysiological annotation pipeline.

5.1 RELATED WORK

User research methods in affective computing

User research methods are usually categorised according to their data source and approach. The data source refers to whether the method measures how its participants act (behavioural) or what they say (attitudinal). On the other hand, the method's approach refers to what type of data is collected (quantitative or qualitative). In game development, although at different stages, virtually all types of user research methods are used. Attitudinal methods (e.g. focus groups, participatory design or desirability studies) are usually applied in earlier development phases, while more concrete (i.e. behavioural and quantitative) methods tend to be used towards the final product delivery deadlines (e.g. beta playtesting periods) (Barakova, Spink, Boris de Ruyter, & Noldus, 2013). Due to the low sophistication of the available techniques, earlier game user research methods focused more heavily on qualitative methods. In this type of method, a participant would play a specific level or level section while a researcher would observe and annotate his session in real-time, a posteriori, or both. These annotations would then be collected for a set of participants representative of the game's target population and later discussed to tune the gaming experience – usually in an iterative design cycle (L. E. Nacke, 2013).

More recent research-based approaches have focused on being able to track game-play events both in real-time and over a larger time frame than that feasible through manual annotation. The most successful of these approaches is Microsoft Game Studios' TRUE (Tracking Real-time User Experience) instrumentation, which allows logging and annotating (i.e. triangulating) interaction events and player feedback. However, despite its achievements, the system fails to take

psychophysiological data into account, which could provide deeper insights into the participant's choices and preferences over time. Complementary to the aforementioned approach, Valve Software has openly announced it is experimenting with biofeedback and psychophysiological user research methods, namely using participants' skin conductance (SC) to measure their relative excitement over several playtesting sessions.

Data Annotation Tools & Frameworks

Given the often times complex design of experimental studies, several tools have been developed to aid in the data synchronization and annotation process. In this section we will discuss the more popular and relevant ones to our needs, whilst comparing them to our proposed tool.

Existing physiological and audio-visual annotation tools

One of the most popular tools for data annotation is Observer XT (Barakova et al., 2013). The tool offers a wide range of data annotation and visualization functionalities, from audio-visual data collection to physiological data visualization and an event logging interface. ANVIL is an also popular, general-purpose tool that allows researchers to track uttered words, head movements, body gestures and other similar inputs on audio-visual data (Maybury & Kipp., 2012). Similarly to the Observer XT, it allows users to augment their annotation with contextual data via an event coding scheme. Being more geared towards speech and body motion analysis, there are also some additional plug-ins that augment its functionality with improved coding schemes (Caldognetto, Poggi, Cosi, Cavicchio, & Merola, 2004).

Current limitations and challenges in physiological data analysis

While the aforementioned tools are able to calculate some statistics from the collected data (number of recorded events, variation and distribution of observed events, event latency, etc.), they remain general-purpose ones that are able to contribute little towards physiological or emotional data analysis per se. In other words, they do not provide a way to interpret the physiological data in any meaningful way (other than these generic statistics) and, more importantly, still require the users to manually code events and emotional responses. These issues have been discussed in (Gunes & Pantic, 2010), where they offer a thorough guide on emotion recognition, which ranges from emotion theories and data modalities to data annotation and interpretation. They have also been discussed in (L. E. Nacke, 2013) from a game analytics context. In their review, Gunes et al. mention several issues with data annotation, one of which being the lack of a standard for emotion recognition, reporting that *"researchers seem to use different levels of intensity when adopting a dimensional affect approach"*. They also refer the issue of inter-observer variability, stating

that *"obtaining high inter-observer agreement is one of the main challenges in affective data annotation, especially when a* offering

dimensional approach is adopted" – an issue found by (Abrilian, Devillers, Buisine, & Martin., 2005) in their creation of a large database of emotionally-coded news clips, where the employed subjective coding technique led to low observer agreement levels (0 to 20% at most at the exception of one emotion) using only 2 observers. The authors conclude that *"(the) development of an easy to use, unambiguous and intuitive annotation scheme remains, however, an important challenge"*.

In a more recent effort towards integrating game-related events with psychophysiological data, (Matias Kivikangas, Nacke, & Ravaja, 2011) have described a system to examine players' physiological responses to game events in post-experiment interviews. While the described system was developed for examining subjective user responses to events, it did not aim at a more objective analysis of the collected data; either by classifying the psychophysiological data in emotional terms or by providing a method for automatically identifying responses to the annotated events.

Standard metrics for affective user experience and human-computer interaction studies

Despite the discussed issues, similar metrics can be found in most affective (Dekker & Champion, 2007; S. W. Gilroy, Cavazza, & Benayoun, 2009; Kuikkaniemi et al., 2010; L. E. Nacke et al., 2011; Wang & Marsella, 2006) and user experience studies (Drachen et al., 2010; Lennart Nacke & Lindley, 2008) involving emotions. This may be largely due to a common agreement on accepted emotion theories and relevant UX metrics, but the wide range of data pre-processing and emotion recognition systems make the objective comparison of experimental results over various sessions or independent studies a challenging task. Additionally, in some cases improper application of known methods may even lead to poor data quality, thus invalidating the study altogether. As such, our proposal aims to be the basis for a standard, embedded emotion recognition system. This would benefit researchers not only by aiding in identifying relevant parts of the recording, but also by standardizing employed methods and obtained results.

It must also be noted that although various significant advances have been made towards emotional detection (Gunes & Pantic, 2010; R. Mandryk & Atkins, 2007; Moreira, 2010), it constitutes a complex problem that would add significant complexity to the development process of any of the discussed tools. In our work, we benefit from being able to easily port our own emotional detection system (PIERS). As mentioned in Section 5:3, we leverage this advantage by directly

incorporating it into our tool (albeit leaving the option for it to be replaced should it be deemed adequate by future users).

Expediting the laborious manual data annotation process

On a more practical sense, the data pre-processing, data interpretation, event coding and emotional reaction extraction phases make up most of the data analysis process. Doing so manually not only requires a substantial larger amount of time, but is also prone to the aforementioned coding errors derived from inter and intra-subject variability. For instance, having multiple researchers analyse and code the same data will likely lead to divergences in coding standards since a certain degree of subjectivity will be involved. Similarly, the same researcher will exhibit variance in his own coding standards which will accumulate with his own fatigue and result in decreasing data annotation quality over time. Being able to expedite this effort would not only result in more trustworthy results, but also enable researchers to do more in less time and reducing the time needed to correct inter and intra person annotation variability or bias artefacts.

Within the context of this thesis we refer to the event coding and emotional reaction extraction phases as emotional reaction triangulation – the process of automatically correlating three arbitrary measures, where the first two (the event and an initial emotional classification) have a causality relation to the third one (the emotional classification posterior to the event).

5.2 REQUIREMENT ANALYSIS AND PARTICIPANTS

Data Annotation Tools & Frameworks

Defining the critical functional requirements for our tool

Despite having our own set of requirements and assumptions, since this was not a tool for private use, it was deemed adequate to gather expert feedback prior to any development efforts. As such, we conducted a series of brainstorming sessions with several other physiological researchers (N=16). Given that psychophysiological research is performed not only by computer scientists, but also by non-technical individuals from the social sciences and psychology fields we took special precautions to ensure all these groups were as equally represented as possible in our study. Ultimately we arrived at the following system requirements:

1. Provide a complete, yet easily interpretable measure of the volunteer's emotional state.

2. A real-time and synchronised view of the volunteer's gaming session from both an audio-visual and psychophysiological perspective.

3. Allow free manipulation of the experiment's rate of time passage (i.e. to quickly scroll through the experience).

4. Allow for a simple and straightforward annotation of relevant events with as few clicks and parameter selection as possible.

5. Present time markers for each of the annotated events and the ability to quickly cycle through and edit them.

6. Enable the user to include subjective, free-form data (comments) for each event, if necessary.

7. Automatically compute which events triggered emotional reactions.

8. Incorporate a save/load feature for resuming the annotation process in relatively large data collections and posterior analysis/verification.

The latter requisite was added in the final stages of our focus group discussions since it was pointed out that studies commonly amass several hours of data (three or more) on a single session. Since these sessions are difficult to reliably annotate in one pass by a single researcher it is not uncommon for multiple researchers to annotate the same session, which requires the annotation process to be resumed a posteriori. Furthermore, several participants stressed the importance of, in addition to the audio-visual and physiological data, being able to import a list describing occurring events (e.g. as outputted by a game log). This list would, in theory, allow the tool to automatically annotate the whole session without any user input.

Catering to the needs of both non-technical and highly specialised audiences

It also became clear that: *1)* the tool should be able to use emotional recognition methods others than our own, and *2)* that their usage should be transparent to the user. Finally, it should be possible to export the identified reactions to a structured output file, so that these could be further examined in greater detail in common third party statistical analysis packages (e.g. R, SAS, SPSS, Weka, etc.). Thus, the following requirements were added to the initial ones:

9. Offer a modular design to accommodate alternative emotional classification or emotional response detection algorithms.

10. Present the ability to not only import the audio-visual and physiological data, but also import a list comprising each of the annotated events (e.g. as outputted by a game engine or logging software) and automatically annotate the whole session without any user input.

11. Transparent emotional classification (i.e. no a priori knowledge needed).

12. The ability to quickly export the identified reactions for later analysis in popular statistical analysis software (e.g. R, SAS or SPSS).

Participants

For the data collection phase, several volunteers (N=22), aged from 22 to 31 years old (M=24.83, SD=2.29) were asked to play two consecutive levels of the survival first-person-shooter (FPS) game Left4Dead 2 (Valve, 2010). This particular game was chosen due to its high event-to-gameplay-length ratio and gameplay diversity, which made it especially suitable for collecting a high number of gameplay-related events in a relatively short time frame.

During the gameplay session, in addition to the physiological metrics necessary for PIERS (SC, facial EMG and HR), we also recorded the gameplay session video using a commercial frame grabber (Fraps - Beepa P/L, Melbourne, Australia, 2007). Upon an initial inspection of the recorded data, it was then processed *a posteriori* using our tool.

Each experimental session was divided in two phases: a first one for obtaining controlled physiological response samples to calibrate the emotional detection system (see Chapter III, Section 3) and a second one where volunteers played the FPS game. In the first phase, participants were shown the same emotional content as in the PIERS study (a relaxing music excerpt, emotionally-charged images from the IAPS library (Lang et al., 2008), and a terror videogame). Participants' responses were then subjectively annotated and used to calibrate PIERS. As aforementioned, the second phase consisted of recording the volunteers' physiological responses whilst playing the FPS game, which were later classified using the emotional detection system and the calibration parameters extracted in the first phase of the experimental session.

5.3 TOOL DEVELOPMENT

Since we wanted to develop a solution that could be used freely regardless of the game engine or stimuli presented in the experimental protocol, we decided to develop a standalone solution. After a brief survey of the available open source libraries and development time cost they imposed, we settled on using C# as the development language. Since our tool is meant to be applicable to a wide range of situations, it requires some parameters to be set: video and physiological data initial

timestamps, emotional classification parameters, types of events and location of video, physiological and annotated event files. For ease of use, we decided to store these values in a simple text configuration file that the application loads at start-up. Furthermore, given the required modular nature (requisite 9), we chose to divide it into various independent components, so that future additions or improvements could be performed in an expeditious manner. These components are, in order of appearance: the emotional recognition module, the event annotation module, and the emotional reaction identification module. Throughout the remainder of this section we will discuss each of the aforementioned modules, how they work and which features they comprise.

Emotion Recognition Module

The first step in the annotation process is determining a simplified, although relevant to our needs, image of the user's current emotional state (ES). As already mentioned in Section 5:1, to achieve this we sought to leverage our previous work, described in chapter IV. To do so, it was necessary to evaluate whether our method (or at least any one of its variants) fulfilled all of the necessary requisites.

Selecting an appropriate emotional state recognition solution

Out of our 12 requisites, 3 are related to the emotional recognition module in some way. Requisite 1 states that we should *"provide a complete, yet easily interpretable measure of the volunteer's emotional state"*. This falls in line with our representation (Russell's circumplex model of emotion) so any version of PIERS fulfils this requirement. Requisite 7 (perhaps the most critical one) asks for the ability to *"automatically compute which events triggered emotional reactions"*. Indirectly, this means that we need to be able to quantify emotional states. Otherwise, measuring an emotional reaction objectively is impossible. Again, using Russell's circumplex model alleviates this issue but means that given the discrete nature of our fusion and ensemble approaches, these are not the best choices for this task. Thus, our choice falls upon our grounded approach, which features continuous classification capabilities. Finally, requisite 11 calls for *"transparent emotional classification"*. Due to the fashion in which the E^2 architecture was conceptualised, this is inherent to PIERS's design so, technically, it is merely a question of porting the code.

Event Annotation Module

Since a considerable proportion of our requisites (55%) were related with how to visualise and annotate the recorded material, we devoted a great deal of attention to the development of the event annotation module. Its function is to address the requirements related to the

annotation functions (2-6 and 10). To fulfil requirements 1 and 2, we decided to combine a custom video player and a time series graph-drawing library (ZedGraph – Champion, Sullivan, 2012). The video capture software automatically logged the starting timestamps of both the gameplay videos and physiological data, which were recorded at 60 and 32 Hz, respectively. The initial timestamps and sampling rates are then given to the tool, which computes the timestamps for each data sample and classifies the physiological data samples in terms of the AV space using the emotion recognition module. Finally, the tool uses the aforementioned timestamps to synchronise the emotional classification and gameplay video streams using a basic linear interpolation process.

Regarding the video player component, it was designed to allow the user to quickly skip through the video using a simple slider or to accelerate the video through a fast-forward and backwards button (requisite 3). The system was later adapted to allow the user to also skip through the data using the emotional classification time plot by clicking on the region of interest to skip to that point in time (Figure 5:1). This was done to improve the tool's usability as sometimes the emotional classification reveals interesting events that might be missed using solely the video.

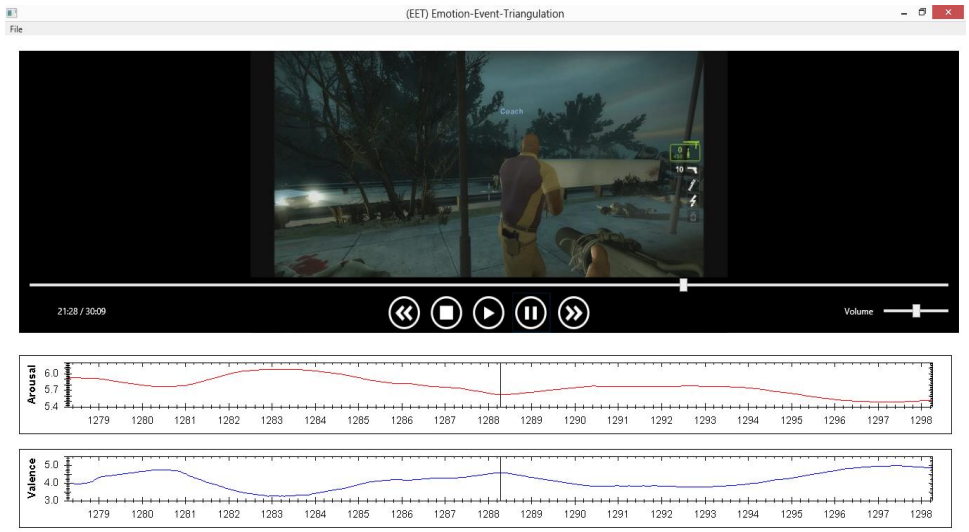


Figure 5:1. A screenshot of the EET tool showing the video player controls (fast backwards, stop, play, pause, fast forward and sound volume respectively), over the emotional classification time plot.

Concerning the event annotation process itself (requisites 4-6), we decided to limit the user input to the barest essentials in both terms of actions and required input. To insert a new event, the user can either perform a right-click on the video player window or right-click on the emotional classification time plot and choose "Add new event" (Figure

5:2). In both cases, this will add a new event at the current time and bring forth a pop-up form where the user can choose which event took place (in case no event file is currently loaded) and any subjective commentary deemed relevant (requisite 6). Finally, the user can access a list of recorded events by, again, right-clicking on the video player or emotional classification time plot and choosing "Edit Events". Double-clicking on any of these events will automatically shift the user the event's timestamp and open its parameterisation window, as if adding the event for the first time.

Making each interactive feature in as few clicks as possible reduces errors, frustration and unnecessary distractions

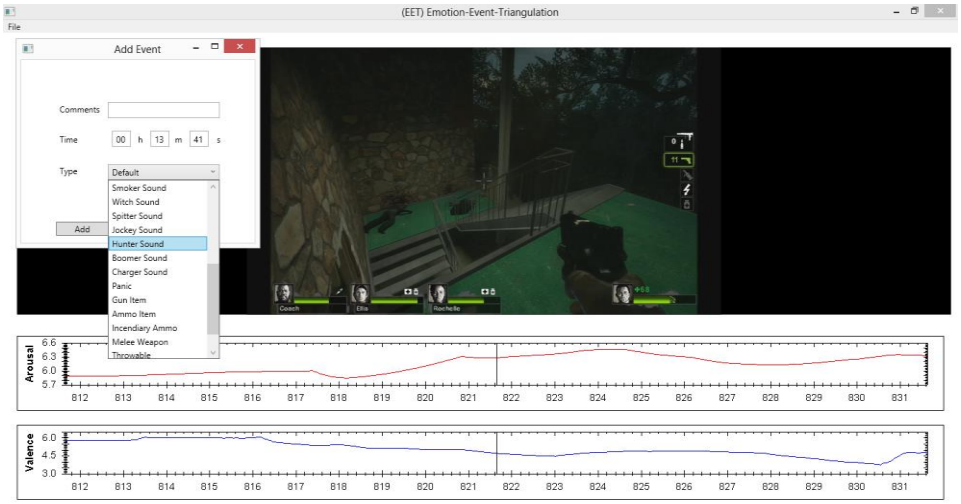


Figure 5:2. The add event window, super-imposed on the EET tool. Although it is editable, the time stamp for the event is automatically filled-in with the current timestamp. The user only needs to choose which event is occurring/going to occur and include any relevant comments (optional).

Finally, attending to requisite 10, we added a feature allowing the user to import a list of previously annotated events. This is done using an optional field in the tool's configuration file, identifying the event list file (see the following sub-section's final paragraph). If the location for such an event list text file is given, the tool automatically parses the file and loads each event. While this was not our case, this feature was added to account for scenarios where event timestamps are automatically generated by other tools. Using this feature, the tool can virtually allow the user to perform the annotation process in a matter of minutes by simply adding the event file name to the configuration file, loading it, and commanding the tool to identify all occurring emotional reactions (see the next section for further details on this process).

Event logs are common in many studies. Being able to interpret a standard file format adds substantial value to our tool

Emotional Response Triangulation Module

The tool's remaining component is the emotional response triangulation module, which is responsible for performing the basic triangulation between the annotated events and the ensuing responses in the AV space (requisite 7). The triangulation process was automated via the following simple local maxima/minima (LMM) detection algorithm. Simply put, the LMM algorithm is a generalization of the '*through-to-peak'* annotation scheme (estimating a peak by using a local – or global – baseline value prior to the stimulus as a comparison), common in physiological recording (Stern et al., 2001). It accepts a time interval relating to an emotional reaction and estimates all local maxima and minima within said interval. A maxima or minima is considered to be any inflexion point with that deviates more than a certain threshold from the mean values preceding the reaction's trigger event timestamp. Since the emotional response triangulation step is the most complex and crucial part of the data annotation process and we intend it to be parameterisable in future versions of the tool, a more formal description is required.

Through-to-peak is an accepted physiological annotation method, making our tool a viable option on many academic studies

Let $c=[c_1,c_2,\ldots,c_n]$ be the continuous, uniformly sampled emotional state classification signal for a dimension of the emotional space (the AV space in our particular case). Furthermore, consider the signal to be smoothed using the unimodal kernel with compact connected support and unit action $w_\Upsilon(t) \geq 0$, and $\Upsilon > 0$ bandwidth parameter through the following process:

$$c_\gamma(t) = w_\gamma(t).c(t) = \int_{-\infty}^{\infty} w_\gamma(t-s)c(s)\,ds$$

The LMM detection process occurs in parallel for both dimensions and is contained in a standalone iteration for each event e_i, within a time interval $\varpi = [max(T(e_{i-1}), T(e_i)-\alpha),\ min(T(e_i)+\beta,\ T(e_{i+1}))]$, where $T: \mathbb{N} \rightarrow \mathbb{R}$ is the mapping function between an event and its corresponding timestamp. Moreover, both α and β are parameterisable event horizon variables (in this thesis α =2 and β =8, as determined by an empirical analysis of the available data). On each iteration of the LMM detection process, the smoothed signal $c_\Upsilon(t)$ is taken and the maximal LMM m is extracted from a set of candidate peaks M:

Parameterization the LMM algorithm plays a big part in its effectiveness

$$M = \left\{ t \in \Omega:\ c_\gamma(t) = \frac{d_{c_\gamma}(t)}{dt} = 0, \left| c_\gamma(t) - c_\gamma(T(e_i)) \right| \geq \varphi \right\}$$

Where φ is a minimum absolute local variability threshold, such that φ = $(\mu_{e_i} + 2\sigma_{e_i})$, with μ_{e_i} and σ_{e_i} denoting the mean and standard deviation values of the considered AV dimension in the processed event's time interval ϖ, respectively.

The maximum 10-second window imposed on ϖ by α and β was specifically designed for this particular study by having in mind: *a)* the response delays of the physiological data used in the emotional classification method (up to 5 seconds for SC), *b)* the time the stimuli usually takes to be perceived – between 1 to 2 seconds due to the lag between the game's telemetry system logging the event and the time it was actually triggered *in-game*, and *c)* the time the emotional response may take to fully manifest itself – in average approximately 1 second, from empirical analysis. They are, however, parameterisable within the tool itself.

While we initially designed the system to identify a single emotional reaction (LMM) subsequent to each event e_i, upon initial analysis of the collected dataset we found that some events had the capability of eliciting multiple (sometimes conflicting) emotional responses. For example, it is fairly common for certain enemies to elicit both low and high valence responses due to the enemy's relation to the gameplay mechanics. Such an example is the Boomer enemy, which is a large, obese character that explodes when shot or within detonation range of the player. As such, it poses both a considerable threat and tactical advantage – if detonated near a group of weaker enemies. It is understandable that when hearing the groan of this enemy type, players felt negative valence (fearing he was close) and then positive valence (after detonating him near a group of enemies). Identifying only the highest (or last) peak in players' emotional reactions would thus potentially discard precious information.

Dealing with composite emotional reactions

To account for this type of emotional responses – which we refer to as composite responses – we adapted the LMM detection algorithm to identify all remaining LMM m_i in M that satisfy the following conditions, instead of the single highest local maxima/minima as we previously did for simple responses:

$$\exists\, m_j, m_i \in M: \left| m_j - m_i \right| \geq \varphi \wedge i > j,$$
$$\nexists\, m_k \in M: \left| m_k - m_j \right| \geq \varphi \wedge \mathrm{j} < k < i$$

Upon extracting this set of LMM values for both arousal and valence, the tool computes the set of corresponding emotional reactions in the AV space by coupling each arousal and valence LMM with their missing coordinate in a tuple set, which is then chronologically ordered. The accuracy results for the LMM detection algorithm can be found in the following section, along with a brief discussion. An illustrative example of the algorithm's output can be found in Figure 5:3.

To fulfil the remaining requisites (8 and 12), this component was also endowed with the ability to export the identified reactions to a

Exporting the identified reactions is a must for performing the traditionally required statistical analysis

structured text file for posterior analysis (requisite 12) and to serialise the entire tool's internal state to a custom *.eet* file extension (requisite 8). The latter allows the annotation process to be resumed or re-analysed in a posterior point in time. Finally, since there is no universally accepted format for physiological data storage, our tool currently accepts the format provided by the BioTrace+ software, which was used in the motivational study. Since statistical analysis software solutions usually accept tab-delimited text files, we chose to export the identified reactions in this format.

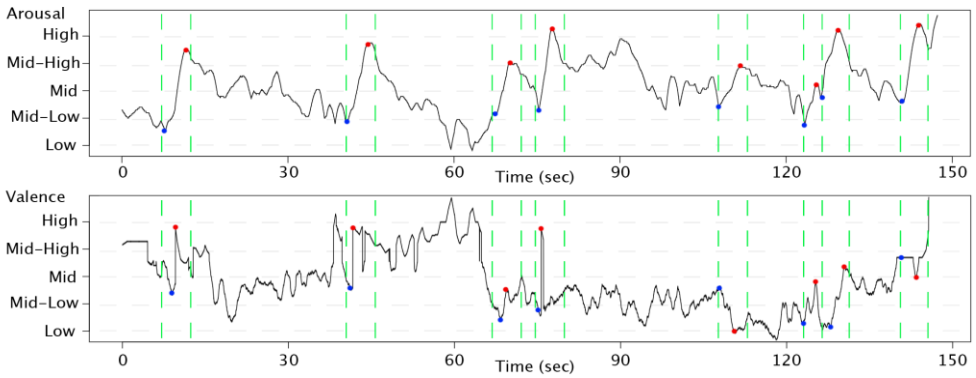


Figure 5:3. Example of the output provided by the peak detection algorithm over a 150-second window. The emotional output was discretised into 5 levels for both arousal and valence for interpretability. Assume that the distance between levels equals the local minimum variability threshold φ. Blue circles denote the timestamp for each logged event; red circles denote the identified local maxima/minima; green dotted lines represent the event's time region (set to 5 seconds for reduced complexity in this example).

5.4 VALIDATION

In order to test the adequacy of the LMM detection algorithm, we decided to compare the results obtained from the automatic detection to a manually validated (ground truth) annotation. While manual annotations are certainly not perfect, we have no data on how many errors occur in a typical annotation so we decided to present results comparing our method to an ideal (manually validated) dataset.

Evaluating the LMM algorithm's performance on real-world data

To evaluate whether the LMM method produced a correct or incorrect response, we used a manual annotation as ground truth. Care was taken to ensure that no errors were introduced in the manual annotation and each reaction was measured independently by each tester. Afterwards they were synchronized to obtain a more robust ground truth. An emotional reaction was considered correct if all peaks

were correctly identified (the highest one for simple responses and all of them for composite ones).

Emotional reactions were obtained from randomly chosen gameplay sessions. Since each gameplay session occurred over a relatively large time frame (μ=37.4, σ=11.4 minutes) and a large number of events were recorded in each one (μ=72.0, σ=28.4), this implied a considerable time effort in manually annotating each session. We thus decided to randomly select six volunteers and use them to validate our algorithm. We could have used random data from each subject, but this would potentially bias the obtained reaction distributions. Instead we chose to focus on analysing a large number of events per participant for a more objective description of the algorithm's performance.

Overall, a total of 430 gameplay-related events were identified and annotated, to which 364 emotional responses were observed – an average of 88.24% event/response ratio, with an 11.52% standard deviation. Out of these 364 identified emotional responses, a considerable minority of them related to simple responses (μ=16.9, σ=8.1), with the remaining 83.1% corresponding to composite responses. This presented an unexpected result that, in our opinion, further justifies the latter enhancement of the LMM detection algorithm. Pertaining the algorithm's accuracy, it revealed adequate performance, as can be observed in Table 5:1. Overall, the algorithm was able to identify emotional responses with a success rate of 93% for simple responses and ~94.5% for composite ones.

Table 5:1. Number of observed emotional responses across all six randomly chosen volunteers and their respective automatic detection accuracy ratings[17].

Volunteer Code	**Number of Responses**		**Detection Accuracy**	
	Simple Responses	*Composite Responses*	*Simple Responses*	*Composite Responses*
A	*17*	*49*	*94.12%*	*91.84%*
B	*11*	*60*	*63.64%*	*98.33%*
C	*8*	*30*	*100%*	*86.67%*
D	*7*	*49*	*100%*	*94.0%*
E	*4*	*78*	*100%*	*97.44%*
F	*12*	*39*	*100%*	*97.40%*

[17] *Simple response detection shows similar performance to composite responses, albeit with a larger standard deviation – probably as a by-product of the smaller sample.*

Volunteer Code	Number of Responses		Detection Accuracy	
	Simple Responses	*Composite Responses*	*Simple Responses*	*Composite Responses*
Total (μ, σ)	*59 (9.8±4.5)*	*305 (50.8±16.8)*	*-- (93±14.6)*	*-- (94.3±4.5)*

A response was considered correctly identified if and only if all LMM were detected. The fact the algorithm presents lower detection accuracy for the simple response category may be justified by both the lower sample population and by its poor performance on volunteer B (whom presented very shallow signal fluctuations, perhaps due to his acquaintance with the game). This led the algorithm to ignore most of his emotional fluctuations, while we acknowledged them in our manual annotation. Despite this, given the rather subjective nature of this task, it remains unclear whether we should have considered these LMM. While this issue could have been solved by simply tuning the algorithm's sensitivity through the φ parameter or by relaxing the constraint placed upon it, we considered a low detection sensitivity to be an adequate trade-off in terms of false negatives versus false positive results for this particular study.

Finally, it is worth mentioning that by considering this particular trade-off, the algorithm did not present any false positive results. In conclusion, in a real life study, researchers should not blindly follow the obtained results, choosing to instead review them and then evaluate whether tuning the φ parameter or relaxing its constraint is a justifiable course of action.

5.5 DISCUSSION AND LIMITATIONS

Estimating potential (time) gains from using our method based on our empirical experience

The system described in this chapter enables game UX researchers to quickly annotate game events and analyse players' emotional responses via their physiologically-classified emotional states. From our own manual annotation process (i.e. not using the tool) we estimated that each event took roughly 30 seconds to annotate. Since using the tool we can simply stop the video when an event occurs, use the GUI to add a game event at the current time and are not required to analyse the emotional state data to determine the player's emotional reaction, this process is shortened to less than 10 seconds. Considering an average of 2 events per minute and a session length of 60 minutes, it would take the following time to process one participant's session:

- Manual annotation (no tool used): 2 hours. One hour for viewing the session video plus another hour for identifying events and analysing the emotional reaction data by hand.
- Semi-automatic annotation (using the tool's emotional reaction extraction algorithm but annotating each event manually in the GUI): 1 hour and 20 minutes. One hour for viewing the session video and an extra 20 minutes taken in adding the observed events through the tool's GUI.
- Automatic annotation (using a game event log): 1-2 minutes. We simply need to import the file containing the game event's timestamps and descriptions and the player's configuration file and ask the tool to export the emotional reactions. Since the user is not required to watch the video to identify when and which events occur, the whole process is virtually automatic.

Notice that we are not taking into account the time taken to review the obtained results since it would take the same amount of time irrespectively of what process was used (manual, semi-automatic or fully automatic). We are also not including the time necessary to format or compute the additional statistics outputted by the tool.

In a final note, while it might seem implausible that an event log is available to enable a fully automatic annotation, this can easily be created in direct observation studies since the researcher can simply record event timestamps while the participant is undergoing the study/treatment. This also applies to game studies as most game engines provide some sort of logging system. We thus estimate that using our tool, researchers will be capable of timesavings of roughly 30%-40% for a semi-automatic annotation (no event log provided) and almost 100% for a fully automatic annotation (event log provided).

Considering typical use cases and the benefits of a standard emotional state recognition and data annotation pipeline

While the automatic emotion recognition is supposed to facilitate the annotation process by guiding users to interesting parts of the recording, since many physiological UX studies also focus on the emotional interpretation of this data, we also expect that by using a standard emotion recognition method our tool will make the obtained results more objectively comparable. The proposed analysis pipeline also aims at reducing the associated workload to the annotation process, while eliminating human subjectivity errors, further contributing towards the standardisation of this study type.

Despite its apparent success, the annotation process is not yet without flaws. Firstly, there is a trade-off between false positive and false negatives in tuning the peak detection's sensibility thresholds. Future work will focus on verifying if the error introduced by this trade-off is not smaller than the one introduced by human error (i.e. inter and

Our tool's current limitations and potential future improvements

intra-subject variability). However, the user can manually correct any automatically obtained results, which eases the issue and still results in a swifter annotation procedure. The second encountered issue relates to the tool's versatility. Since it is logistically impossible to integrate it with every existing game engine or application, user's must always rely on either manually building log files or import existing ones – for which parsers may not always be readily available. There are also some limitations regarding gameplay annotation of events that have longer durations (e.g. a monster chase). These events do raise some interesting questions in terms of automatic response estimation as they prompt a tonic fluctuation on players' emotional state rather than a phasic one. Since this requires a deeper analysis of the emotional signal's structure and would imply a much more complex validation, we have refrained from implementing this feature in this version of the tool.

5.6 SUMMARY

Current UX research methods are unable to perform in-game evaluations without disrupting – and thus potentially contaminating – the gameplay experience. Moreover, emotional state classification methods are difficult to integrate in these studies due to their complex development nature and technical skillset. The methodology presented in this chapter has the potential to contribute to a wider accessibility of emotional response studies by, not only easing the aforementioned issues, but also by removing the necessity of developing standalone emotional state detection systems – which in itself contributes to a standardisation and comparability of the annotation process.

In principle the tool fulfils all of the established requirements while retaining a generalizable approach – a feature not widely adopted in earlier related work. This versatility is dictated by the tool's independence towards the input data and by the emotional recognition module's modular design (which nonetheless can be exchanged by another implementation). Furthermore, the tool does not limit the data analysis process to its own limited capabilities, as it allows the user to export the detected emotional responses for further scrutiny or modelling.

Contributions present in this chapter and how they enable our progress towards affective player reaction models

Within the more general narrative of this thesis, we have proposed, implemented and validated an objective method for extracting emotional responses from players' physiologically interpreted emotional states. Recalling our thesis objectives, this corresponds to part A of objective V: "*Propose a grounded methodology to… automatically associate the emotional reactions to the eliciting interaction events*".

We are now thus in position to focus our efforts on modelling players' emotional reactions to Vanish's game events. In chapter VII we will attempt to model these reactions using, firstly, a purely feature-driven (black box) approach. Driven by the lack of a continuous output, difficulty in model updating and sparse nature of our dataset, we then shift towards a more domain-knowledge – but still data-driven – clustering method that enables both continuous output, seamless incorporation of new training input and handling of sparse data.

REFERENCES FOR CHAPTER V

Abrilian, S., Devillers, L., Buisine, S., & Martin., J. C. (2005). EmoTV1: Annotation of Real- life Emotions for the Specification of Multimodal Affective Interfaces. In *HCI International*. Las Vegas, USA.

Barakova, E. I., Spink, A. S., Boris de Ruyter, L., & Noldus, P. J. J. (2013). Trends in measuring human behavior and interaction. *Personal and Ubiquitous Computing*, *17*(1), 1–2. doi:10.1007/s00779-011-0478-x

Caldognetto, E. M., Poggi, I., Cosi, P., Cavicchio, F., & Merola, G. (2004). Multimodal Score: an ANVIL™ Based Annotation Scheme for Multimodal Audio-Video Analysis. In *International Conference on Language Resources and Evaluation Workshop on Multimodal Corpora* (pp. 29–33).

Dekker, A., & Champion, E. (2007). Please biofeed the zombies: enhancing the gameplay and display of a horror game using biofeedback. In *Situated Play, Proceedings of the Digital Games Research Association (DiGRA) Conference* (pp. 550–558).

Drachen, A., Nacke, L. E., Yannakakis, G., & Pedersen, A. L. (2010). Correlation between Heart Rate, Electrodermal Activity and Player Experience in First-Person Shooter Games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games* (pp. 49–54). ACM.

Gilroy, S. W., Cavazza, M., & Benayoun, M. (2009). Using affective trajectories to describe states of flow in interactive art. In *Proceedings of the International Conference on Advances in Computer Enterntainment Technology - ACE '09* (p. 165). New York, New York, USA: ACM Press. doi:10.1145/1690388.1690416

Gunes, H., & Pantic, M. (2010). Automatic, Dimensional and Continuous Emotion Recognition. *International Journal of Synthetic Emotions*, *1*(1), 68–99.

Hazlett, R. (2006). Measuring Emotional Valence during Interactive Experiences : Boys at Video Game Play. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 1023–1026).

Kuikkaniemi, K., Laitinen, T., & Turpeinen, M. (2010). The influence of implicit and explicit biofeedback in first-person shooter games. In *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 859–868).

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS).*

Leon, E., Clarke, G., Callaghan, V., & Sepulveda, F. (2007). A user-independent real-time emotion recognition system for software agents in domestic environments. *Engineering Applications of Artificial Intelligence*, *20*(3), 337–345. doi:10.1016/j.engappai.2006.06.001

Mandryk, R., & Atkins, M. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*, *65*(4), 329–347. doi:10.1016/j.ijhcs.2006.11.011

Matias Kivikangas, J., Nacke, L., & Ravaja, N. (2011). Developing a triangulation system for digital game events, observational video, and psychophysiological data to study emotional responses to a virtual character. *Entertainment Computing*, *2*(1), 11–16. doi:10.1016/j.entcom.2011.03.006

Maybury, M. T., & Kipp., M. (2012). Multimedia Annotation, Querying and Analysis in ANVIL. In *Multimedia Information Extraction: Advances in Video, Audio, and Imagery Analysis for Search, Data Mining, Surveillance, and Authoring*. doi:10.1002/9781118219546

Moreira, V. H. V. G. (2010). *BioStories Geração de Conteúdos Multimédia Dinâmicos Mediante Informação Biométrica da Audiência*.

Nacke, L. E. (2013). An introduction to physiological player metrics for evaluating games. In *Game Analytics* (pp. 585–619). Springer London.

Nacke, L. E., Kalyn, M., Lough, C., & Mandryk, R. L. (2011). Biofeedback Game Design : Using Direct and Indirect Physiological Control to Enhance Game Interaction. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (pp. 103–112). ACM.

Nacke, L. E., Stellmach, S., & Lindley, C. a. (2010). Electroencephalographic Assessment of Player Experience: A Pilot Study in Affective Ludology. *Simulation & Gaming*. doi:10.1177/1046878110378140

Nacke, L., & Lindley, C. A. (2008). Boredom, Immersion, Flow - A Pilot Study Investigating Player Experience. In *Conference on Game and Entertainment Technologies*.

Russel, J. A. (1980). A Circumplex Model of Affect. *Personality and Social Psychology*, *39*(6), 1161–1178.

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). *Psychophysiological recording* (2nd ed.). New York: Oxford University Press.

Wang, N., & Marsella, S. (2006). Introducing EVG: An Emotion Evoking Game. In *Intelligent Virtual Agents* (pp. 282–291).

Chapter VI

AFFECTIVE REACTION MODELS

MODELLING PLAYERS' EMOTIONAL REACTIONS VIA PHYSIOLOGICAL INPUT

OUTLINE

While individual emotional responses extracted from players' emotional state signal offer some insights into how they perceive certain game events, they fall short of a complete description on how these game events are affected by their pre-stimulus emotional state or how they evolve over time. Furthermore, current approaches to game design improvements employ time-consuming gameplay testing processes, which rely on highly subjective feedback from a target audience.

In this chapter we explore two distinct approaches at modelling players' emotional reactions in a comprehensive manner. Both approaches are intended as input-agnostic and generalizable across games and game genres, as well as other forms of digital media.

Our first approach relies on extracting a set of features from players' emotional responses to game events, which are then used to drive a feature selection algorithm that outputs an optimal feature set based on a hill-climbing method. Based on the optimal feature sets found by three feature selection algorithms (best first, sequential feature selection and genetic search), the collected features are used to create computational models of players' emotional reactions on the arousal and valence dimensions of emotion, using several machine learning algorithms.

Our second approach also attempts to extrapolate the causal relations between changes in players' emotional states and recorded game events through a hierarchical clustering model of their approximated (regressed) reaction models. These clusters can later be used to infer individual player models via fuzzy cluster membership vectors, providing us a solution with continuous output, human-interpretable models and sparse data resistant properties.

We expect this work to benefit game developers by accelerating the affective playtesting process through the offline simulation of players' reactions to game design adaptations and contribute towards individually-tailored affective gaming as we will see in chapter VII.

Emotionally adaptive games are one of the holy grails of modern affective game research. However, the current state of the art on affective gaming relies on static game adaptation mechanics that assume a fixed emotional reaction from players every time. In Chapter IV we studied these games and various types of biofeedback mechanics

programmed to react to players' emotional states according to a fixed set of game design / adaptation rules. We saw that while providing statistical and empirical evidence of their effectiveness, these mechanics are not enough to create a completely adaptive emotional experience. Furthermore, even if in a purely testing scenario, commercial game design is based on game design optimizations via typical beta-testing procedures, which falls short of ideal both in the level design and long-term gameplay experience fronts.

Potential benefits of affective reaction profiles and a light overview on our approach towards them

Along with improving the existing gameplay experience, predicting players' emotional reactions can give game developers a powerful tool to more finely-tune the original experience before releasing a game and accelerate the game design process - ultimately resulting in not only improved, but also more quickly produced titles.

Although we present an initial feature-driven approach, this chapter's core contribution is a method for modelling player's emotional reactions to game stimuli through fuzzy memberships to player clusters. These clusters are obtained through a bootstrapped hierarchical clustering algorithm and validated using approximately unbiased (AU) and bootstrapped probabilities (BP). This clustering approach initially attempts to approximate individual players' emotional response functions to game events based on a set of observed emotional reactions. Individual player model pairs are then compared to find groups of players that share similar reactions, through the hierarchical clustering algorithm. Using the clusters' more stable models, we compute players' fuzzy membership vectors to the found clusters, based on a simple distance function.

6.1 RELATED WORK

Player Modelling

A focused introduction to player modelling for affective gaming

Player modelling has been a popular topic in Artificial Intelligence, mainly because of the substantial advantage that knowing in advance how a player (usually an opponent) might choose to act provides in terms of strategy definition.

Most recent approaches have taken this concept and attempted to use them for modelling player experience (Yannakakis & Togelius, 2011). The most straightforward way to do this is through player's questionnaire responses. Although this process may create very accurate models (Shaker, Yannakakis, & Togelius, 2010), the considerable presence of experimental noise (derived from human error in self-judgment, memory, etc.) and the intrusiveness of the method can

lead to difficulty in analysing the data. The work by (Tognetti, Garbarino, Bonarini, & Matteucci, 2010) shows how self-reports can be successfully used to capture aspects of player experience. Similarly, Shaker et al.'s work on this area models players' experience on an emotional basis (Shaker et al., 2010) using crowd-sourced data regarding player actions and level design features. Although these models provide some insight to players' affective state, the models' low granularity cannot capture all the nuances of human affect.

Analysing player modelling from traditional, subjective surveys to modern, objective physiologic data

A less intrusive approach relies on using gameplay data (e.g. how many times a certain action was performed) in an attempt to build these models (Etheredge, Lopes, & Bidarra, 2013). The main assumption is that player actions and preferences are linked to player experience, making it possible to infer players' emotional states by studying their interaction patterns (C. & H., 2009; Gratch & Marsella, 2005). This approach is the least intrusive one, thus becoming a candid possibility for real world usage. However as stated in (Yannakakis & Togelius, 2011), the models are often based on several strong assumptions that relate player experience to gameplay actions and preferences, resulting in a low-resolution, often unsatisfactory or over simplistic, model of playing experience and its affective component.

Prior work on physiological player modelling has achieved promising results in predicting subjective player experience reports using simple physiological metrics (Lankes, Hochleitner, Hochleitner, & Lehner, 2012; Martinez, Garbarino, & Yannakakis, 2011; Vachiratamporn, Legaspi, Moriyama, & Numao, 2013). Game narratives have also been shown to be dynamically adaptable in response to players' physiological state (S. Gilroy & Porteous, 2012), proving this technology can be applied to various facets of the gaming experience. We hypothesize that the success obtained by these approaches may be due to their usage of a more objective and continuous data source that arguably lowers the amount of noise in the data labelling process and providing a rich data source.

While, as in the aforementioned work, we employ physiological metrics, our method is input agnostic, abstracting the emotional data and game events as an n-dimensional waveform and class labels, respectively (see Sections 6:2 and 6:3). As in the previous chapters, the emotional detection method employed is PIERS's grounded approach, as described in Chapters I and III.

6.2 DATA COLLECTION & FEATURE EXTRACTION

Revisiting and characterising the gameplay data collected during our static IBF study with Vanish

Recalling our case study from Chapter IV, Vanish is a survival horror videogame where players must navigate a network of procedurally generated maze-like tunnels to locate a set of key items, before being allowed to escape. During gameplay, the player must avoid a creature that continuously stalks him. Several visual and audio events (e.g. lights failing, steam or water pipes bursting or the creature distant cries) also occur, in order to keep the player engaged. These events, along with creature encounters, deaths and locating new items are automatically logged by the game and constitute the set of considered game events.

One of the logistic challenges when dealing with physiological data is acquiring enough data from unbiased (gameplay) sessions. The dataset used in this chapter was extracted from the 72 gameplay sessions that the 24 participants from our study in Chapter IV allowed us to record. In total, they comprise over 30 hours of annotated gameplay (Figures 5:3 and 6:1) and physiological data obtained through a complex, Latin-Square balanced experimental protocol. As per the defined experimental protocol (see Section 4:3), each player underwent a brief tutorial to learn the game controls and was left alone in the room to avoid tainting the collected data.

In this section we start by describing the feature set extracted from the emotional responses and used by the feature-based approach. In order to properly describe our clustering approach, we then formalise our data sources – the physiological data, interpreted as emotional states through PIERS and the ensuing emotional reactions.

Emotional States

A formal definition of emotional states over time and within the context of emotional responses

As stated, our aim is to model players' individual emotional response functions to a set of game events $G=\{g_1,...,g_k\}$, given an initial (henceforth referred to as prior) emotional state λ_p, such that $\lambda_p \in \Lambda$, the set of considered emotional states.

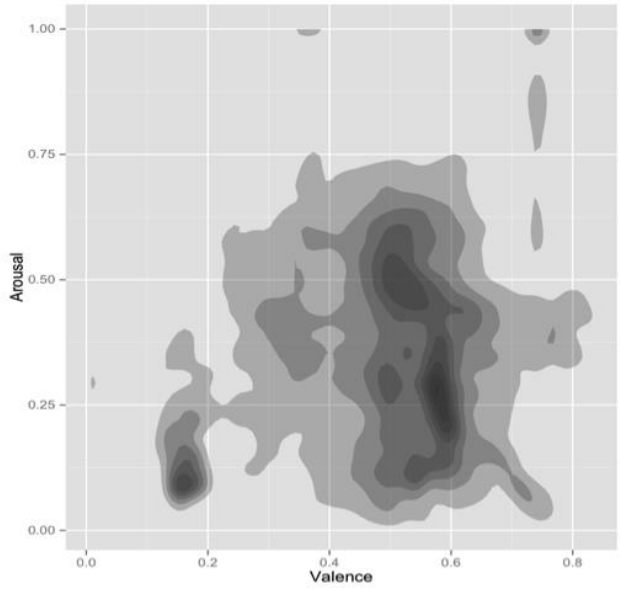
While neither of the methods presented in this chapter are reliant on physiological data or emotional detection system – any continuous n-dimensional waveform is acceptable (see definitions 1 and 2) – we will be using PIERS's grounded implementation as our only data source within the scope of this thesis. Despite this, the experimental details such as data collection and feature extraction process are still documented for reproduction purposes.

Definition 1 (*Emotional State*). An emotional state $\lambda \in \Lambda$ is defined as a n-tuple $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n)$, where each element represents a dimension of the considered emotional theory representing the emotional state.

In our case $\lambda = (A, V)$, where:

- A is the observed value in the arousal space, with $A \in \mathbb{R} \mid A \in [-1, 1]$
- V is the observed value in the valence space, with $V \in \mathbb{R} \mid V \in [-1, 1]$.

Definition 2 (*Emotional State Waveform*). The emotional state waveform W is defined as the tuple $W = (w_A, w_V)$, where w_A and w_V are the continuous, uniformly sampled emotional state classifications for arousal and valence, respectively and are of the form $w_x=[w_{x1}, w_{x2},\ldots,w_{xn}]$.



Surprisingly, for a horror game, Vanish was able to elicit a considerably wide emotional spectrum

Figure 6:1. Density plot of the normalised elicited emotional spectrum over all 72 gameplay sessions.

Emotional Reaction Extraction

The processed physiological data produced two emotional state waveforms w_A and w_V with a 1:1 mapping for each study participant over a wide emotional spectrum (see Figure 6:1 for an overall view of the elicited emotional spectrum). The waveforms, along with the game's event log metadata, were then synchronized and used to extract player's emotional reactions (see Definition 3 – which is a summarised recap of our earlier formalisation of composite responses).

Revisiting the concept of (composite) emotional reactions

Definition 3 (*Emotional Reaction*). Consider a specific instance of a game event g_i and its corresponding timestamp g_i^t on the emotional state waveform w. Consider also the time interval T associated with this specific event g_i, such that $T_g = [max(g_{i-1}^t, g_i^t - \alpha), min(g_i^t + \beta, g_{i+1}^t)]$, where both α and β are parameterisable event horizon variables (as mentioned in the previous chapter, in this thesis α =2 and β =8). Briefly recalling the formalisation in chapter V, we define an emotional reaction to a game event g_i and the pair of local maxima or minima (m, m') of each emotional state dimension, taken from the prior $T_m = [max(g_{i-1}^t, g_i^t - \alpha), g_i^t]$ and subsequent $T_{m'} = [g_i^t, min(g_i^t + \beta, g_{i+1}^t)]$ time intervals that exhibit the highest distance between them. Both m and m' are extracted from a set of candidate peaks M such that:

$$M = \begin{Bmatrix} t \in T: w(t) = \frac{d_w(t)}{dt} = 0, \\ |w(t) - w(g_i^t)| \geq \varphi \end{Bmatrix} \qquad \text{(Eq. 6:1)}$$

Where φ is a minimum absolute local variability threshold, such that $\varphi = (\mu_{e_i} + 2\sigma_{e_i})$, with μ_{e_i} and σ_{e_i} denoting the mean and standard deviation values of the considered AV dimension in the considered time interval, respectively.

Thus, an emotional reaction r is defined by the triggering game event instance g_i, the prior emotional state λ_p and the response emotional state λ_r formed by the conjunction of the local maxima or minima pairs (m, m') of each emotional state dimension: $r = (\lambda_p, g_i, \lambda_r)$.

Sanity checks and data quality assurance are a critical part of model development

Despite Vanish automatically generating the event timestamps – which were synchronised with the AV ratings – we took care to manually inspect the obtained emotional responses to make sure no annotation noise was introduced into our dataset.

Unfortunately, upon an initial analysis, two subjects did not have their reactions recorded due to a logging system malfunction, and another one was not included due to insufficient records. Furthermore, since some events occurred very sparsely (e.g. dying), they were not considered, as their inclusion would artificially inflate the classifiers' performance. Overall, three event types were discarded. After this filtering process, the used dataset registered over 1,160 emotional reactions, distributed among 12 unique gameplay events.

The extracted reactions also underwent a post-processing cycle, which consisted in manually removing or inputting outliers to the distribution's mean. These outliers were mostly null reactions, which corresponded to events that were generated by the game engine but, for

some reason, were not triggered/seen by the player or occurred simultaneously with other game events. In the case of simultaneous events, the reactions were ignored since, at this point in time, we were neither interested or in possession of sufficient data to accurately model the compounding effects of simultaneous game events.

Habituation effects were a concerning issue that could significantly undermine the size of our dataset and should be taken into account in future studies

We also hypothesized that players might exhibit light to moderate habituation effects to the game events, which would seem a reasonable assumption over time. However, it would appear that due to the experiment's relatively short timeframe (~40 minutes), these effects were not visible in our dataset. Nevertheless, this is a real possibility in longer studies, so an exponential averaging function (or similar weighting mechanism) should be applied over consecutive conflicting reactions - i.e. reactions with the same initial emotional state and triggering event that exhibit significantly[18] different response emotional states.

Additional Features Extracted for the Feature-Driven Approach

For the feature-driven approach, a total of ten extra features were extracted for each emotional reaction: five related to arousal and another five pertaining valence levels. Both valence and arousal share the same feature extraction process. Onwards from the gameplay event timestamp, the following features are created:

- $E\{r\}$: Initial value, calculated as the average of the maximum and minimum values registered ($avg(max\{r\} + min\{r\})$). This is the base feature extracted by our triangulation tool
- $\mu\{r\}$: Mean of the signal
- $\sigma\{r\}$: Standard Deviation of the signal
- $M\{r\}$: Maximum Value of the signal
- $m\{r\}$: Minimum Value of the signal
- D_h : Absolute time period between minimum and maximum value ($D_h = | t^{h}_{max}\{r\} - t^{h}_{min}\{r\} |$)
- $h_{in}\{r\}$, $h_{out}\{r\}$ and $h_{ev}\{r\}$: Auxiliary features denoting the reaction's beginning, ending and event timestamps.

The delta values of the reactions (ΔA / ΔV) are calculated as per the

[18] One possible definition for a significantly different response emotional state would be as an outlier, when compared with the previous response states. However, simply exponentially averaging new responses would more elegantly resolve this issue.

LMM algorithm explained in Chapter V and, as the initial value feature ($E\{r\}$), are the same for both the feature-driven and clustering-based approaches.

6.3 A FEATURE-DRIVEN (BLACKBOX) APPROACH

Determining Optimal Feature Sets

Feature selection is an expensive process but should not be necessary in subsequent iterations to new game genres as human physiologic response processes are static across the emotional spectrum

Feature selection is a common concern in machine learning problems since identifying the most relevant feature combinations to maximize classifier performance is not always a trivial task. Carefully selecting an appropriate feature set results in improved model generalisation, reduced overfitting, shorter model training times, and higher model interpretability - which given our applicational focus on allowing game designers to semi-automatically improve player experience is of significant relevance.

Therefore, we apply several feature selection algorithms (FSA) to identify the feature subsets that maximize model accuracy without performing an exhaustive search on all possible feature combinations. In this section we examine the optimal feature subsets (OFS) generated by the *n best individual feature selection* (BestFirst), *genetic search* (GSearch), *sequential forward selection* (SFS), and *random search* (RSearch) algorithms. Notice that since these algorithms all rely on some variant of *hill climbing*, they cannot guarantee anything more than a locally optimal feature subset.

Table 6:1. Number of features selected per feature selection algorithm, for arousal and valence reaction prediction.

	BestFirst	**GSearch**	**SFS**	**RSearch**
ΔA	2.12	3.09	2.12	4.02
ΔV	2.14	3.07	2.14	4.12

Table 6:2. Feature selection algorithm errors (RMSE).

	ΔA		***ΔV***	
	Mean	SD	Mean	SD
BestFirst	0.2843	0.2105	0.3730	0.3184
GSearch	0.2872	0.2136	0.3768	0.3200
SFS	0.2843	0.2105	0.3730	0.3185
RSearch	0.2884	0.2124	0.3842	0.3304

Overall, it seems that virtually all feature selection algorithms are able to achieve very similar prediction errors (see Table 6:1). However, a closer analysis of each FSA's average optimal feature subset size (Table 6:2) reveals that despite the similar performance, the average number of selected features per OFS varies considerably between FSA. In general, the BestFirst and SFS are the most efficient algorithms, leading with the least average number of features used (~2.13), followed by the GSearch (~3.08) and RSearh (~4.07), in second and third place respectively. This, in conjunction with a simple analysis of the top 10 features chosen by each FSA (Table 6:3), clearly indicates that most of the predictive power resides in a small subset of extracted features. This could hint that the difficulty lies not on selecting the optimal feature set, but possibly in modelling the classification function in the underlying data instead. It also suggests that while effective, similar computational intelligence techniques might not always be ideal for tackling this type of issue, in particular when using physiological data. For a more thorough discussion of this notion, please refer to the discussion section (Section 6.6).

Table 6:3. Top 10 selected features per feature selection algorithm, for both arousal and valence reaction prediction.

BestFirst		**GSearch**		**SFS**		**RSearch**	
$\boldsymbol{\Delta_A}$	$\boldsymbol{\Delta_V}$	$\boldsymbol{\Delta_A}$	$\boldsymbol{\Delta_V}$	$\boldsymbol{\Delta_A}$	$\boldsymbol{\Delta_V}$	$\boldsymbol{\Delta_A}$	$\boldsymbol{\Delta_V}$
$\sigma\{r_A\}$	$\sigma\{r_V\}$	$\sigma\{r_A\}$	$\sigma\{r_V\}$	$\sigma\{r_A\}$	$\sigma\{r_V\}$	$\sigma\{r_A\}$	$\sigma\{r_A\}$
36%	33%	37%	42%	36%	33%	51%	56%
$E\{r_A\}$	$\sigma\{r_A\}$	$\sigma\{r_V\}$	$\sigma\{r_A\}$	$E\{r_A\}$	$\sigma\{r_A\}$	$\sigma\{r_V\}$	$\sigma\{r_V\}$
31%	31%	37%	33%	31%	31%	45%	45%
$\sigma\{r_V\}$	$E\{r_A\}$	D_h^V	D_h^V	$\sigma\{r_V\}$	$E\{r_A\}$	D_h^A	D_h^V
30%	28%	32%	31%	30%	28%	41%	44%
D_h^A	$E\{r_V\}$	D_h^A	$m\{r_V\}$	D_h^A	D_h^V	D_h^V	D_h^A
22%	23%	30%	29%	21%	22%	41%	40%
D_h^V	D_h^V	$E\{r_A\}$	D_h^A	D_h^V	$E\{r_V\}$	$\mu\{r_A\}$	$E\{r_V\}$
21%	21%	29%	28%	21%	21%	36%	37%
$M\{r_V\}$	D_h^A	$M\{r_A\}$	$E\{r_A\}$	$M\{r_V\}$	D_h^A	$M\{r_A\}$	$m\{r_V\}$
14%	19%	23%	*26*%	14%	19%	30%	36%
$M\{r_A\}$	$M\{r_V\}$	$m\{r_V\}$	$E\{r_V\}$	$\mathrm{M}\{r_A\}$	$M\{r_V\}$	$m\{r_V\}$	$M\{r_A\}$
12%	14%	23%	*25*%	12%	14%	30%	30%
$E\{r_V\}$	$m\{r_V\}$	$\mu\{r_A\}$	$M\{r_A\}$	$m\{r_V\}$	$m\{r_V\}$	$\mathrm{m}\{r_A\}$	$\mu\{r_A\}$
12%	12%	23%	*23*%	12%	14%	27%	28%
$\mu\{r_A\}$	$M\{r_A\}$	$m\{r_A\}$	$M\{r_V\}$	$\mu\{r_A\}$	$M\{r_A\}$	$E\{r_V\}$	$\mu\{r_V\}$
9%	12%	22%	*20*%	9%	12%	27%	27%
$m\{r_A\}$	$\mu\{r_V\}$	$E\{r_V\}$	$\mu\{r_V\}$	$E\{r_V\}$	$\mu\{r_V\}$	$E\{r_A\}$	$m\{r_A\}$
9%	11%	20%	*18*%	9%	11%	26%	25%

Emotional Reaction Models based on Optimal Feature Sets

In order to build the players' affective reaction models, we chose the Linear Regression (LR), Multilayer Perceptron (MLP), and M5 Model Trees and Rules (M5P) classifiers. Since the available number of observations for each event varies significantly (e.g. the number of "death" events is an order of magnitude lower than the "scare" events) two different evaluation modes are presented. In order of preference the methods are: 10-fold cross-validation, 3-fold cross validation.

Neural networks seem to be the most adequate classifier for this task, perhaps due to their "plasticity"

Overall, it would seem that our previous suspicions – that the greatest difficulty would be in modelling the classification function itself – were correct. Table 6:4 shows that the MLP classifier was much more efficient than the LR and M5P classifiers, achieving ~35% and 25% better results in arousal and valence reaction prediction, respectively. In contrast, the LR and M5P classifiers showed similar results.

Finally, since we empirically noticed some players reacted dissimilarly to different game events, we were interested in seeing whether allowing the player models to use different classifiers for each game event would significantly impact the observed prediction accuracy. These results are presented in the last line of Table 6:4 and indicate no significant improvements can be seen in contrast to the MLP classifier, which acts as a testimony to its overall performance.

Table 6:4. RMS error ratings for arousal and valence reaction predictions.

	ΔA		ΔV	
	Mean	SD	Mean	SD
LR	0.3114	0.0754	0.4010	0.1388
M5P	0.2959	0.0671	0.3962	0.1271
MLP	0.2389	0.0933	0.3238	0.1393
Best	0.2309	0.0802	0.3121	0.1234

Discussion / Conclusion

By combining emotional state reactions inferred from physiological data with a computational intelligence approach we were able to predict players' emotional reactions with a reasonably low error rate. However, despite the relatively large sample for an experimental study involving physiological data, we believe that further increasing the sample size would result in better results. Due to the obvious logistic constraints involved in our study, it was also impossible to validate our approach on

different games and game genres, which implies the need for further studies.

Regarding the approach itself, it is not without its intrinsic limitations. First and foremost, this type of computational intelligence approach's main issue is that it is unable to capture player models at different levels (e.g. player types). It is possible to create models to predict the reactions of a specific player group, but player groups must be identified and labelled a priori. Furthermore, on its extreme, the models are faced with a difficult trade-off; either generalise over an entire population, which may be less than ideal when we would like to optimize the gameplay experience specifically for one particular player, or predict how an individual player would react. The latter on, although ideal for adaptive affective gaming, is often unviable due to the difficulty in eliciting the sheer amount of necessary emotional reactions to properly map the feature space; lest we risk overfitting the model itself. Moreover, the only way to account for new emotional reactions (and thus the unavoidable habituation effects to previously experienced game events) is to reprocess the collected dataset, extract new feature sets and then rebuild the player models. Although this might be computationally cheap, there may be scaling issues and it requires a somewhat sophisticated automated processing pipeline.

The main challenge in predicting individual reactions is collecting enough data from a single player in a timely manner

Finally, neural networks' black-box design, along with their tendency to make unstable predictions when extrapolating on unseen regions of the feature space leads us to the conclusion that a more stable, interpretable approach would be desirable. An approach that would simplify the model updating process and, more importantly, provide a continuous output while being robust in the face of sparse data would in fact be ideal. Given the high predictive power of the E{r} and σ{r} features, we believe that a more grounded approach using either of these features could be used to build (computationally) simpler player models that could then be clustered based on a similarity coefficient. On the following section we will thus focus our efforts on improving the models' usability and robustness using a hierarchical bootstrapped clustering approach.

Neural networks are volatile, difficult to scrutinize and do not offer continuous data output

6.4 CLUSTERING AFFECTIVE PLAYER MODELS

As we have previously discussed, our aim is to model players' individual emotional response functions to a set of game events, given a prior emotional state. More formally, these models should obey the generic function **Φ** (Eq. 7:2):

$$\Phi : \Lambda XG \rightarrow \vec{v} \mid \sum_{i=0}^{len(\Lambda)} \overrightarrow{v_i} = 1 \ \wedge \overrightarrow{w_i} \in [0,1] \qquad \text{(Eq. 6:2)}$$

Where function Φ receives a prior emotional state λ_p and a game event g, and outputs a weight vector $\vec{v}$ that contains the probabilities of observing a transition to each of the possible emotional states Λ ($|\Lambda| = |\vec{v}|$), if the considered game event g is performed at the prior emotional state λ_p.

In this section we describe how the extracted reactions are used to approximate players' emotional reaction functions to game events using a more grounded approach. We start by approximating players' responses individually using their prior emotional state (equivalent to the E{r} feature) and a multi-dimensional regression matrix (Definition 4). We then describe how these models were used to cluster player pairs based on a similarity matrix using a bootstrapped hierarchical clustering method.

Approximated Player Models

Individual player models provide a grounded, gross approximation of player's general reactions

Definition 4 (*Approximated Player Model*). Let $\mathcal{R} = (r_1, r_2, \ldots, r_n)$ be the set of emotional reactions extracted for a single player p and $\mathcal{R}_g = (r_1^g, r_2^g, \ldots, r_k^g)$ be the sub-set of emotional reactions to a particular game event g, such that $\mathcal{R}_g \subset \mathcal{R}$. An approximated player model (APM) can be defined as the approximation (i.e. conditional expectation) function of the independent variables – the player's response emotional state λ_r – in regards to the dependent variables – the player's prior emotional state λ_p and triggering game event g. Given that we logically assume players' reactions to be independent to game events (i.e. each game event influences the player differently), game events are regressed separately, which also reduces the overall model complexity and training time. Moreover, since no assumptions can be made on the form of the players' underlying emotional reaction function, as no theoretical framework exists, a non-linear model, as proposed in previous works would seem advisable (Martinez et al., 2011; Shaker et al., 2010). Thus, the player's reaction to a specific game event can be modelled as a multivariate non-linear regression function based on his emotional state prior to the triggering game event:

Regression models enable numeric predictions and generate continuous prediction surfaces

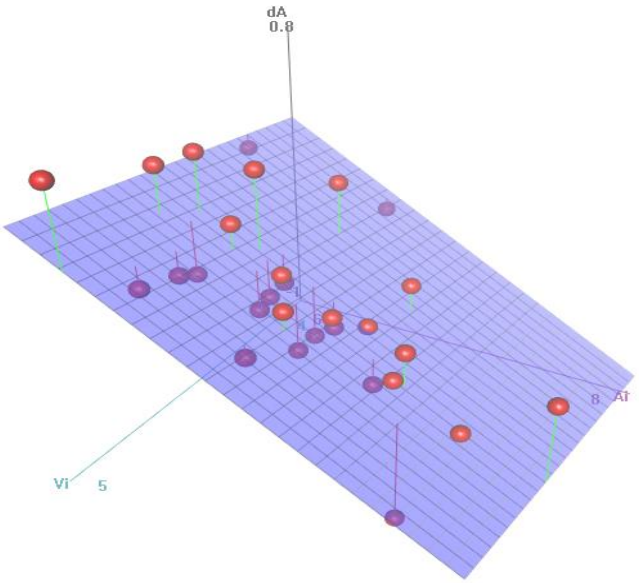


Figure 6:2. . Illustrative example of a 3D cross-section of a player model fitted using a linear regression. For interpretability, the image shows one of the model's 16 dimensions for predicting the arousal component of the reaction (ΔA) to game events.

$$Y = \beta_0 + \beta_{11}X_1 + \cdots + \beta_{1n}X_1^n + \cdots + \beta_{p1}X_p + \ldots + \beta_{pn}X_p^n + \varepsilon_i \quad \text{(Eq. 6:3)}$$

Reactions to game events are modelled individually, keeping the feature space dimensionality broken up

Where $X = (X_1, \ldots, X_p)$ denote the dimensions of the prior emotional state, n denotes the regression's polynomial order (in this thesis p=2 and n=3) and Y represents the predicted value for the dimension of the response emotional state being modelled. A player's APM is thus defined as a set of k=nm regression surfaces that represent his emotional reaction functions to each of the individual n game events over the m dimensions of the AV space (4). In our work, n=16 and m=2.

$$APM = M_{n,m} = \begin{bmatrix} Y_{11} & \cdots & Y_{1m} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{nm} \end{bmatrix} \quad \text{(Eq. 6:4)}$$

Small sample sizes for some events may skew the model considerably so restrains must be set in place

Due to their high expressiveness, polynomial regression models are known to easily *overfit* the available data (albeit to a much lesser degree than other popular machine learning techniques such as the neural networks used in our earlier approach). Thus, upon a visual analysis of our dataset, we decided to adopt a supervised stepwise regression scheme capped at third order polynomials. Another common issue with machine learning techniques is the rapidly increasing degree of uncertainty when extrapolating (i.e. the error involved in making

predictions outside of the training dataset's value range quickly rises). To counter this issue we applied a tapering function to each built model, restricting it from predicting values outside $\mu \pm 2\sigma$ of the dependent variable. This prevents the model from predicting illogical response values due to simple extrapolation errors. An illustrative example of a 3D cross-section of a player model can be seen in Figure 6:2.

Distance Matrices

In order to cluster player's according to the similarity of their emotional reactions, we require an inter APM distance metric (Definition 5).

Definition 5 (*Inter APM Distance*). Let M and M' be two approximated player models, whose distance is given by:

Comparing the shape of the regression surfaces between player pairs provides an objective metric for player similarity

$$\delta(M, M') = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d(M_{ij}, M'_{ij})}{|M|} \qquad \text{(Eq. 6:5)}$$

Where $d(Y, Y')$ represents the distance between two regression surfaces, which, in turn, is given by:

$$d(Y, Y') = \frac{\sum_{i=1}^{b\Delta^{-1}}\sum_{j=1}^{b\Delta^{-1}} |e^{Y(i\Delta,j\Delta)} - Y'(i\Delta,j\Delta)|}{|b\Delta^{-1}|^2} \qquad \text{(Eq. 6:6)}$$

Where Δ is the sampling granularity of the continuous regression surface generated by the model Y (Δ = 0.1 in this thesis). Although various metrics can be employed in this step (e.g. Euclidean, Manhattan), not to mention numerous transforms (e.g. sigmoid, logarithmic), an exponential function allowed us to easily differentiate between increasingly dissimilar models in a non-linear fashion appropriate for the clustering approach.

Put differently, the inter APM distance represents the average similarity of two players across equivalent game events and emotional response dimensions. Notice that since we have no supporting evidence that any game event poses a higher influence on the game's affective experience, all elements of the APM matrices were equally weighted (Eq. 7:5).

Upon computing the inter APM distances for each approximated player model pair, they are organized into a distance matrix that is fed to the clustering algorithm.

Bootstrapped Hierarchical Clusters

To find player clusters, we applied a hierarchical clustering algorithm to the distance matrix obtained from computing the inter APM distance for each player pair. In order to obtain an initial estimate on the clusters' significance, we applied a multi-scale bootstrap resampling process, which allowed us to obtain Approximately Unbiased (AU) *p*-values for each identified cluster (see Figure 6:3) (Suzuki & Shimodaira, 2006). These values represent the confidence that a particular cluster is supported by the data, as opposed to a random sampling error effect. Formally, for a cluster with an AU *p*-value of *x* the null-hypothesis "*the cluster does not exist*" is rejected with a significance level *s* = 100-*x*. As such, high AU values lead us to the belief that the cluster would be stably observed with an increasing number of observations.

Player clusters deliver a more robust, statistically significant estimate on player responses and can be used to classify new players with little data

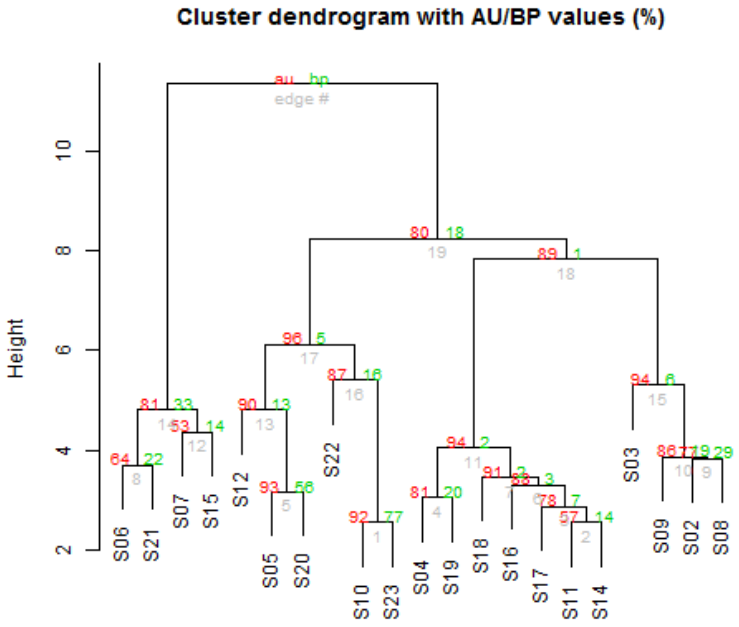


Figure 6:3. Hierarchical clusters with AU p-values (red/left) and BP (green/right).

Cluster Analysis & Validation

While the AU *p*-values provided by the bootstrap resampling process indicate that the created clusters are well supported by the data (see Figure 6:3), they do not offer a tangible proposition as to the optimal number of clusters.

Common approaches are to cut the tree at the largest links or to manually analyse the created clusters to determine the most relevant ones (Etheredge et al., 2013; Holmgård, Togelius, & Yannakakis, 2013). However, more objective and less domain knowledge dependant techniques exist. Two of the most widely accepted ones are the Sum of

Squared Error (SSE) and Dispersion coefficients (Eq. 6:7 - 6:8), which measure the within cluster cohesion and between cluster separation, respectively.

$$SSE = \sum_i \sum_{\in Ci} d(M, Ci)^2 \quad \text{(Eq. 6:7)}$$

$$Dispersion = \sum_i |Ci| d(C, Ci)^2 \quad \text{(Eq. 6:8)}$$

Since cluster cohesion (inverse SSE) and dispersion naturally increase as more clusters are created, a stopping criterion is required. Popular choices include zero-crossings and inflexion points, as these denote critical points where further splitting the data results in diminishing returns. We chose the inflexion points as both curves had noticeable inflexion points at the same cluster number (Figure 6:4).

Determining cluster number and significant is as much of an art as science so we approached it from multiple points of view to confirm our assumptions

The created clusters were also manually cross-matched with the demographics data reported by the players; genre preference (whether players liked horror games or not), gamer type (casual or hardcore) and sex. Results were encouraging and showed clear divergences between clusters. For example, we found that *C1* contained only male hardcore players, while *C4* was made up of mainly softcore players that disliked horror games. However, this does not mean that, for example, *C1* contains all male hardcore players, which suggests the collected demographics are able to roughly outline the player characteristics but additional information would be necessary to characterize the found player groups. Unfortunately, due to opportunity limitations this analysis is out of the scope of this thesis as it would require an additional, wider demographic study.

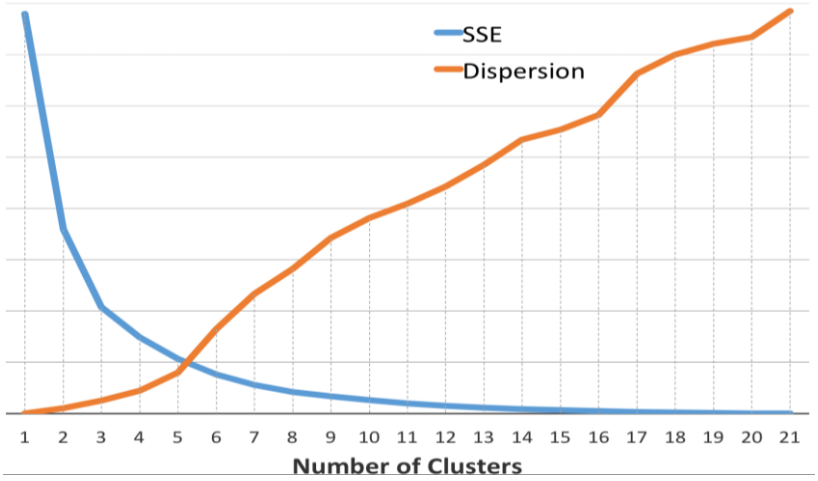


Figure 6:4. SSE and dispersion values for increasing cluster counts.

6.5 EXTRAPOLATING INDIVIDUAL PLAYER MODELS THROUGH FUZZY MEMBERSHIPS

While the identified player clusters allow us to examine how particular player types would react under any given game state, they also remove the possibility of individually predicting players' reactions. Using players' approximated models instead could easily solve this issue, but doing so would not allow us to benefit from their increased robustness due to higher amount of data used to train them. It would also defeat their usefulness in modelling new players by reusing the hierarchical model itself to classify them on early stages of data acquisition.

Clusters provide high-level representations of player groups.

To address these issues, we model each player according to a fuzzy membership vector, expressed through his relative distance to each cluster (see Definition 6). This is a common approach in soft clustering methods (e.g. fuzzy c-means) to allow a certain degree of uncertainty when making predictions and also to differentiate between members of the same cluster. Since we wanted our model to be capable of representing unseen players without rebuilding the cluster models, we decided to compute the membership function outside the clustering process. Figure 6:5 presents the obtained player-cluster membership vectors.

Like a child, each player derives his uniqueness by inheriting features from many of them

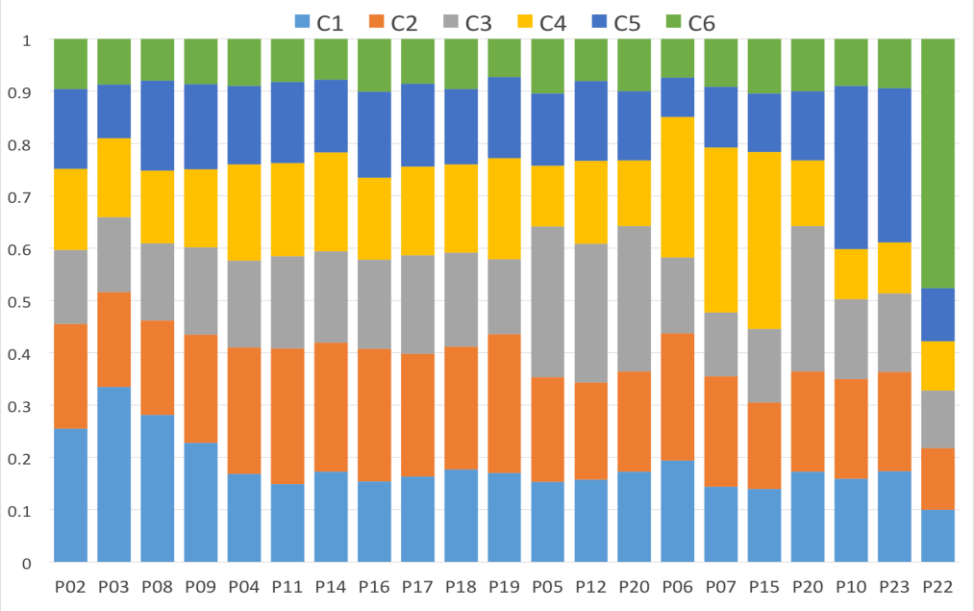


Figure 6:5. Individual player membership vectors. Columns represents players and shaded areas the individual cluster membership probability for clusters C1 to C6.

Definition 6 (*Fuzzy Player Model*). Let $C=\{C_1,C_2,...,C_n\}$ be the set of clusters as identified by the hierarchical clustering algorithm. Also, consider a cluster C to be the average of the APM models $\{M_1,M_2,...,M_n\}$ associated with it. A player with an APM M belongs to each cluster C in inverse proportion to his distance to the cluster model (Eq. 6:9) after being normalised over all existing clusters (Eq. 6:10):

$$I(M,C_i) = \frac{\sum_{i=1}^{|C|} d(M,C_i)}{d(M,C)} \qquad \text{(Eq. 6:9)}$$

$$\rho_i = \frac{I(M,C_i)}{\sum_{j=1}^{|C|} I(M,C_i)} \qquad \text{(Eq. 6:10)}$$

Thus, the final player model P can be expressed in terms of its fuzzy cluster membership vector $P = [\rho_1,\rho_2,\dots,\rho_n] \mid n = |C|$ that determines the weight each individual cluster response should be given when predicting the player's emotional response.

6.6 DISCUSSION

The results presented by the clustering approach reveal that besides the selected clusters, most other possible clusters are heavily supported by the data. Besides indicating that we were able to successfully differentiate between players based on their emotional reaction functions, it also suggests that re-interpreting the clustering structure at different levels might lead to clusters possibly representing very dissimilar player groups.

No longer bound to nominal, non-volatile predictions, black-box models or needing to choose between individual or group models

In comparison with our initial (feature-driven) approach, we have been able to improve on its limitations on various points. Firstly, we are no longer bound to its trade-off between generalizing over an entire population and focusing on a specific individual, thus risking overfitting. Secondly, we have gained the ability to make numeric predictions, backed with a non-volatile and continuous model of user affect. The human-interpretable individual player models are also of practical utility for manual tweaking / detection of data outliers and useful for more subjective evaluations of player behaviour in other experimental or behavioural studies.

The hierarchical clustering approach can also infer player responses based on fewer data

The added robustness and stability brought by the clustering method combined with our fuzzy membership approach also contributes to the method's ability in coping with new players not included in the original dataset. For example, it is possible to assign a player to a specific cluster (or set of clusters to be more precise) based on a small, sparse set of emotional reactions and, based on the more complete

cluster models extrapolate – with a certain degree of certainty – how he would react to unmapped parts of the feature space on his approximated model.

Another potential application would be in artificially diversifying study populations. Since by clustering we obtain a generic idea of how each player type (cluster) would react, it is easy and viable to produce new synthetic players by generating membership vectors that follow the clusters' natural distribution parameters. These synthetic players can then be used in a deep simulation (e.g. a Monte Carlo-based engine) to perform extensive automated tests on potential gameplay adaptations. In fact, given the logistic limitations in performing extensive human studies we adopt this technique for our final validation experiments, as we will see in Chapter VII.

Synthetic player models allow us to extensively test game adaptations free of logistic issues such as habituation effects, fatigue or increasing subjectivity

Besides this, the relatively low computational cost involved in building these models also means that re-computing them *on-the-fly* is feasible as more data becomes available, meaning they are able to adapt as players experience more of the game. This should help alleviate our previous approach's potential scaling issues. While further validation is still required for other game genres, as we model player reactions to game events individually, the method should scale well to other games with hundreds of event types (e.g. MMORPGs).

As with our previous approach, the biggest limitation is the study size (in both number of collected samples and game genres tested). We expect that aided by the formalisation of our method, future studies will contribute to this issue, validating and suggesting new improvements to our approach.

6.7 SUMMARY

The proposed models are not only useful for affective games, but can also be used for offline game optimization procedures by using the player models to estimate how players would react to specific game configurations, and only then performing live tests with human players. This can provide game designers with a powerful tool to accelerate the game design and testing process, as well as a means to identify where their game excels or falls short of achieving the desired affective experience.

Chapter summary and limitations

Despite the limitations of the presented approaches, both are reliable and generic methods for modelling players' affective reactions on virtually any game genre. We would however like to note that games

Presented contributions and immediate applicability

that do not meaningfully elicit any physiological alterations on the player would result in, albeit correct, null response models. We would also like to point out that while we focus on physiologically interpreted emotional state data, this is not a requirement of our method and any numerical representation of the player's emotional state will suffice. The same is true for the arousal and valence dimensions of emotion. As mentioned in the previous section, although we do not know how well the updating process might scale on games featuring much larger event sets (e.g. role playing games with hundreds of event types), given the size of the feature space, the theoretical model holds.

One of the most immediate contributions posed by our work is the potential to both accelerate and increase the objectivity of the (affective) game design process. For example, besides using the models to drive the adaptation of the gameplay experience automatically in real-time, the affective player models can be used during *playtesting* phases to inform game designers on which stimuli are most effective at eliciting a set of target emotional states. In terms of our thesis' objectives, in this chapter we have addressed part B of objective V: "*(To) propose a grounded methodology to… compile the observed emotional reactions into players' affective reaction models"*.

Next steps towards validating the models' emotional elicitation capabilities

In the following chapter we will be using our models as the key-driver component for a symbolic simulator that allows semi-automatic tuning of game parameters according to players' reactions. This symbolic simulator will run an abstracted version of Vanish and, through the use of players' models, allow the identification of optimal game parameters for a target affective experience. Our analysis will focus on two main points: *1)* whether the models can be used *offline* to determine the best game configurations parameters to elicit a specific emotional state (i.e. to a scenario where the game does not adapt at all, but is optimally calibrated for a specific experience) and *2)* whether the models can be used to elicit a specific emotion or emotional pattern by doing real-time adaptations to the gameplay experience. The latter makes up our final thesis objective (VII) and we expect these results will constitute an (initial) validation on the real-world usage of physiological emotional reaction models in adaptive affective gaming technologies.

REFERENCES FOR CHAPTER VI

C., C., & H., M. (2009). Empirically Building and Evaluating a Probabilistic Model of User Affect. *User Modeling and User-Adapted Interaction*, *19*, 267–303.

Etheredge, M., Lopes, R., & Bidarra, R. (2013). A Generic Method for Classification of Player Behavior. In *IDPv2 2013 - Workshop on Artificial Intelligence in the Game Design Process*.

Gilroy, S., & Porteous, J. (2012). Exploring passive user interaction for adaptive narratives. *Intelligent User Interaction*, 119–128.

Gratch, J., & Marsella, S. (2005). Evaluating a Computational Model of Emotion. In *Autonomous Agents and Multi-Agent Systems* (pp. 23–43).

Holmgård, C., Togelius, J., & Yannakakis, G. (2013). Decision Making Styles as Deviation from Rational Action A Super Mario Case Study. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Lankes, M., Hochleitner, W., Hochleitner, C., & Lehner, N. (2012). Control vs. complexity in games: comparing arousal in 2D game prototypes. In *Proceedings of the 4th International Conference on Fun and Games* (pp. 101–104).

Martinez, H. P., Garbarino, M., & Yannakakis, G. N. (2011). Generic Physiological Features as Predictors of Player Experience. In *Proceedings of the 2011 Affective Computing and Intelligent Interaction Conference* (pp. 267–276).

Shaker, N., Yannakakis, G., & Togelius, J. (2010). Towards automatic personalized content generation for platform games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (pp. 63–68).

Suzuki, R., & Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, *22*(12), 1540–1542.

Tognetti, S., Garbarino, M., Bonarini, A., & Matteucci, M. (2010). Modeling Enjoyment Preference from Physiological Responses in a Car Racing Game. In *IEEE Conference on Computational Intelligence and Games* (pp. 321–328).

Vachiratamporn, V., Legaspi, R., Moriyama, K., & Numao, M. (2013). Towards the Design of Affective Survival Horror Games: An

Investigation on Player Affect. In *Affective Computing and Intelligent Interaction (ACII)* (pp. 576–581).

Yannakakis, G. N., & Togelius, J. (2011). Experience-driven Procedural Content Generation. *Transactions on Affective Computing*, *2*(3), 147–161.

Chapter VII

MODEL VALIDATION VIA SIMULATED PLAYOUTS

EVALUATING PLAYERS' REACTION MODELS VIA SIMULATED PLAYOUTS

OUTLINE

So far, through this thesis, we have proven that: 1) it is possible to reliably measure players' emotional states in real-time, 2) biofeedback mechanisms, even of the static form, significantly affect players' emotional states and gameplay experience, and 3) creating affective reaction models from players' semi-automatically inferred emotional reactions is feasible.

However, despite having successfully built statistically significant affective reaction models for our player demographic, we have yet to confirm our most critical thesis hypothesis: that these models are capable of, in theory, eliciting a pre-determined set of emotional states.

In this chapter we focus on this final hypothesis by using the previously generated player models to drive a symbolic representation of Vanish. In this final experiment, our focus will be two-fold: 1) to determine whether the player models can be used in an offline fashion to determine optimal gameplay parameters for a non-biofeedback experience (i.e. used for optimising the game prior to its release), and 2) to assess if, by consulting the model's predictions during the gameplay experience, we can influence players' emotional states according to a set of target emotional states and patterns (i.e. emotionally regulate the gameplay experience in real-time).

We start by providing a brief introduction to the work presented in this chapter and then move towards a description of the simulator's conceptual design. However, our main focus for this chapter will be on the aforementioned experiments and on assessing the fitness and implications of the obtained results.

In Chapter VI we saw that the created hierarchical clusters cleared both our objective and subjective analysis. From an objective perspective, the Approximately-Unbiased estimates were all statistically significant for all created clusters and from a subjective standpoint, players' attributed to specific clusters roughly shared similar traits, such as gamer proficiency. However, the fact that we were able to create models of players' emotional reactions is still insufficient to accurately assess whether biofeedback affective experiences are capable of eliciting specific, target emotional states.

The challenge that is put before us is thus to assess in which scenarios are these models of considerable usefulness. As stated in

Chapter I, we originally envisioned two main application types for our work:

A compromise between real-world applicable and state of the art breaking results

1. In a more immediate application on gameplay testing and affective computing studies to determine which gaming/experimental conditions more closely elicit the desired emotional experience
2. As a full-blown emotionally adaptive system capable of using players' historic data to produce tailored affective experiences.

Evidently, the latter is much more ambitious but it is also of questionable feasibility considering the current market for physiologic gaming. As most advances are made in incremental steps, we believe it is important for this technology to gain traction before commercial affective gaming solutions become a reality. This rationale was present, for example, in our decision to open-source our emotional reaction triangulation tool, presented in Chapter V. Likewise, for this thesis to be of interest to a wider (potentially more practical or industry-focused) audience, it is essential that our results present actionable opportunities. As such, we have chosen to assess the models' emotional elicitation capabilities on both these application scenarios.

Having made this decision, we are now faced with a complex dilemma: how to perform this validation? Performing a user study as the one presented in Chapter IV would perhaps seem ideal but is riddled with logistic issues. While most of these are theoretically surpassable, virtually every one of them is, in practice, insurmountable. The following are the most relevant obstacles we encountered when considering a user study:

Logistic issues with new live human studies

1. Habituation effects: While we didn't encounter habituation effects on our original Vanish dataset, it is natural that as players get accustomed with the game events, their reactions to them change over time.
2. Increased sample size: As the number of tested conditions increases, so must the sample size. Having in mind the available human resources, the complex experimental protocol and high data processing cost, this alone substantially undermines the feasibility of such a study.
3. Order effects: Since this study would require us to test a large number of gaming conditions, we would be faced with either introducing order effects into our results of further increasing the sample size.
4. Psychological fatigue: Asking players to go through the game for extended periods of time would not only cause habituation

effects, it would also stress them (physically and mentally) and most likely influence their responses as the game would no longer have any ludic component.

5. Participant availability: Sourcing participants willing to commit 2 to 3 hours of their time to our study was moderately difficult (~25% withdrawal rate). Doing so for a study at least an order of magnitude above would prove of questionable feasibility, at least without some form of substantial monetary compensation.

Faced with these limitations, we turned out attention towards a popular alternative in agent systems research; to simulate our case study. This allows us to create an environment where we can test our models free of the aforementioned contamination factors. Also, given the high number of configurations allowed by Vanish's procedural engine (>1,000), and the fact that we wished to test the emotional impact of, at least a representative sample of them, this approach presented itself as ideal. A simulation approach also enables us to not only perform a study orders of magnitude above what a user study would realistically allow, but also to repeat the same playouts any number of times to make sure our results are stable.

A simulation based approach as an alternative to human studies

Despite all of these advantages, this approach is not without its drawbacks. Firstly, the most evident one is that some aspects of the simulation may be incomplete (e.g. the interplay between a light source and the creature's placement, or some aspect of the gameplay experience that was not modelled but that influenced the player). While we found no significant emotional reactions outside of the gameplay events, it is a fact that the simulation is a proxy for reality so its results will always offer a trade-off between an extensive analysis and players' actual experience.

Limitations and additional challenges or a simulation based approach

Secondly, using a simulation approach requires the development of a dedicated solution that replicates Vanish's gameplay experience as closely as possible. Luckily, this is not an issue since we have full access to the game's core components and can easily re-implement the procedural generation engine rules and even override them based on our models' predictions as necessary.

Thirdly, given the large amount of intended simulated playouts (see Section 7:2), we cannot use the game's rendering engine. In fact, even running it in "headless mode"[19] is not an option as the physics, lighting and AI components (among others) would take up too much processing time, making the simulation process too costly for the scale we have

[19] Headless mode refers to an execution mode where the program runs all of the logic and control code but does not perform any rendering of the created content. It is often used to perform automated unit and functional tests.

defined[20]. The solution to this issue naturally resides in reducing the amount and complexity of the components necessary for the simulation, which led us to adopt a symbolic approach (see Section 7:1).

Necessary optimizations due to computational constraints and required additional player modelling

Finally, while we have modelled players' emotional responses to the game's events, we are missing two components from their overall behaviour: their gameplay style (how they interact with the game world) and their emotional decay rates (the evolution of their emotional states in the absence of game events).

Having substantiated our choices for both the experimental conditions and validation approach, we will now focus on describing the development of our symbolic simulator, along with its simulation modes / flow (Section 7:1). In the second part of this chapter we will present our experimental protocol (Section 7:2), the obtained results (Section 7:3) and their implications (Section 7:4).

7.1 A PROCEDURAL SYMBOLIC SIMULATION ENVIRONMENT

As we have previously discussed, performing the simulations using Vanish's native codebase was not a viable option. The solution was to find an alternative that abstracted away as much of the game's computational cost as possible, without losing its meaning or fidelity. In this sub-section we will discuss how we achieved this using a symbolic approach to represent the game's state, which optimization were necessary and, finally, how we addressed the issues pertaining players' gameplay styles and emotional decay rates.

Given that the simulator is a high-level abstraction of Vanish's procedural world generation and gameplay mechanics – which have already been presented in great detail in Chapter IV – an in-depth discussion of how the game's playout is simulated is not warranted. In this section our focus lies primarily in describing which abstractions and code optimizations were implemented in order to achieve acceptable simulation run times (see footnote 21).

From a workflow perspective the simulation process starts by specifying a series of configuration parameters on which the simulator should run. These parameters are placed in a CSV file that is read by the simulator on start-up and it contains the following fields:

[20] Even if running the game in headless mode achieved a 10x increase over actual gameplay, testing 10,000 game variants, each lasting 10 minutes would take up approximately 1 week of uninterrupted simulation time. Running the simulations once per player would take almost half a year, not accounting for hardware failures, bugs or repeated runs for output stability checking.

- Initial emotional state: The players' initial arousal and valence levels (individually defined). Default value is 5.0 for both (i.e. a neutral emotional state).
- Base (atmosphere-induced) emotional state: The emotional state towards which players drift to on the absence of game events. As the game as an "atmospheric emotional charge", it is natural that players are not on a neutral emotional state when playing the game. Being a horror game, this is usually a mildly low valence and slightly elevated arousal level (a state within the tension zone of Russell's AV space). This value was set using players' own gameplay data from Chapter IV's study.
- Emotional state decay: At what rate should the player's emotional state drift towards the base emotional state (measured in AV points per second). Upon empirical analysis, default values were set at 0.025 and 0.05 for arousal and valence, respectively.
- Game event parameter vector: This vector tells the simulator which game events are available (all, by default), and what are the possible configurations. Since the simulator is intended to go through a large set of game configurations, this vector defines which values each parameter (e.g. environment events, creature encounters, hallucinations, etc.) can take. This is achieved by defining a minimum and maximum bound for how often each event can occur (measured in seconds), as well as an associated incrementing/decrementing step. With these triples, the simulator can then generate all possible game configuration variants in runtime, without the need to hardcode these into the system. It also makes limiting the simulation to a specific configuration subspace much more user-friendly. To determine the range on which to define the configuration subspace we naturally performed small deviations over the Vanish's tested vanilla configuration, thus preserving the game's overall identity.
- Simulation repetition count: How many times each game configuration should be run, per player. Given the pseudo-random nature of Vanish's gameplay, we considered it good practice to run the same simulation multiple times to reduce result variance. Default value for this parameter is 50 for test runs. Our experiments were run with it set at 100.
- Simulation playtime: Maximum duration for each simulation. This pertains to how long of a gameplay experience should be simulated, not how long each simulation has of CPU time. Default value is 600 seconds, which allowed for a complete game playthrough. Most simulations ended before this mark, either

Describing the simulation parameters and their influence on playouts

due to the player succeeding in escaping or dying/becoming insane.

- Time step: The simulation's time step. Default value was 10 seconds, meaning each simulation was completed within 60 iterations of the simulator's main loop.
- Intrusiveness: How often should the player's model be consulted to adapt the gameplay experience (in simulation time steps). The default value was 3, with the minimum being 1 and maximum being 6.
- Target emotional state: The emotional state that the player should have during the gameplay session. In the case of dynamic states (see Section 7:2), these are a sequence of AV tuples that must form a closed pattern (i.e. the initial and final states must be the same).
- Model evaluation depth: How many game events can occur within the same simulation time step. Since more than one event can occur within a 10 second time frame we can allow the player model to search which combination of game events more closely positions the player to the target emotional state. However, to avoid artificially boosting our results it is necessary to limit this search to the number of game events that could naturally occur under non-adaptive version of Vanish. Default value for this, given a time step of 10, is 3.

All of these parameters can be defined through a simple graphic user interface. Once the user is satisfied with them, he can run the simulator by selecting one of our two experiences (Figure 7:1) and the simulator will make use of the appropriate parameters.

After it is finished running, the simulator outputs a log file with the results for the experiment. The log file details all of the player's emotional states for each step of each simulation, as well as every event that was triggered. A debug version of the log was also used for testing purposes that outputted information on how long specific instructions took and how specific aspects of the simulation (e.g. game world generation) were computed for optimization and debugging purposes, respectively.

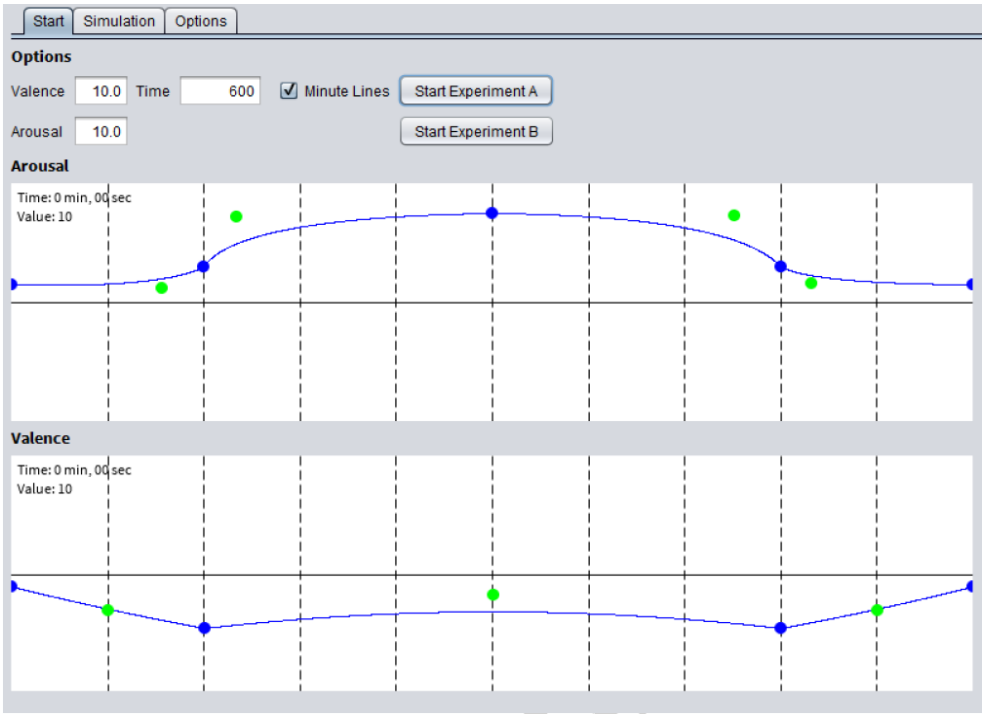


Figure 7:1. The GUI designed to parameterise and run the simulator. Special attention was given to the definition of emotional patterns, which can be drawn using Bezier curves and are then translated as a series of AV tuples in the simulator's configuration file.

Default simulator configuration and produced output

Overall, we tested approximately 4,000 possible game variants by regularly sampling from the game event parameter space allowed by Vanish's procedural engine. Each of these game variants ran for a maximum simulation time of 600 seconds on 10 second time steps and was repeated 100 times, per player. Roughly, this totalled nearly 600 million simulation cycles for one experiment, which took 8 hours to run on an i7 2.7GHz quad-core CPU. The production version of the log file had approximately 33 GB of pure text data for both experiments and had to be processed offline by a dedicated script.

Simulating Game World Components via Layering

As we have previously discussed, most of our computational cost cutting exercise is focused on eliminating the rendering and computer graphics related overhead in our effort to achieve a symbolic representation of the game's state that still retains a high degree of fidelity.

To achieve this, we separated the gamestate's representation into three layers. Layer 1 has the representation of the game world; what blocks currently exist, what assets/game events they can afford, and the players' current position in the game world. Using Vanish's game design

Symbolic simulation layers and their adaptation from Vanish's native codebase and design grammar

grammar, presented in Chapter IV (recall figure 4:4), this was virtually an immediate port from its game world generation rules and symbols. Layer 2 defines the creature's position in the game world, its positioning within each block (hidden or visible) and its predisposition towards the player (i.e. which AI state it's in – see Chapter IV, Section 4). Given that the creature is dynamically spawned on a specific block and the transition rules between its AI states are easily translated into a basic finite state automaton, this layer posed no significant challenges. Finally, layer 3 represents the positioning of game events. Despite this layer being tied to layer 1, the game world, the way in which the simulator changes and adapts each one is different and would require additional computations were we to keep them closely coupled. This layer also uses Vanish's design grammar to define which events are valid on the existing blocks but abstracts away all of the anchoring and variable trigger timing details. These three layers define a game state; a valid configuration of the game world/experience in a specific point in time. For illustrative purposes, Figure 7.2 presents a visual representation of layer 1 on top of an existing Vanish game level.

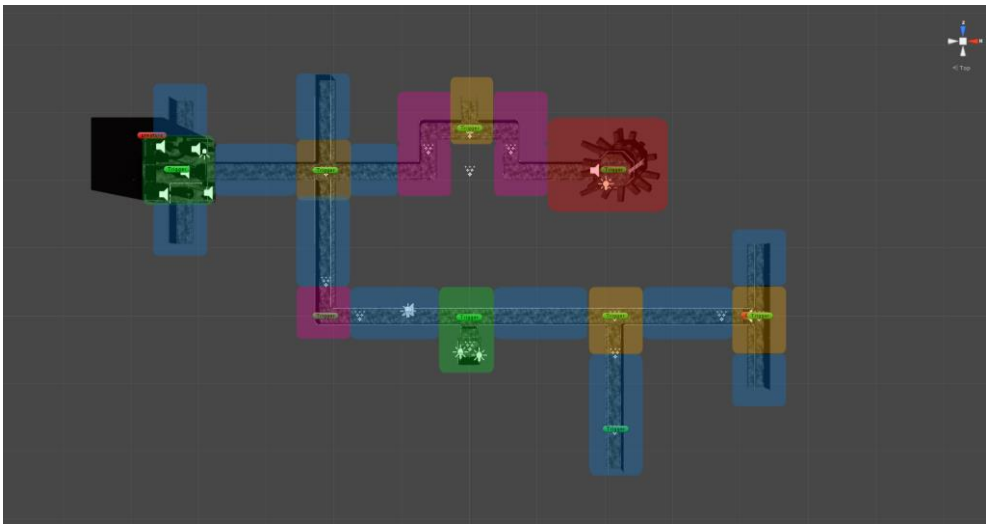


Figure 7:2. Visual representation of Vanish's game world symbolic abstraction. Each differently coloured section is a symbol/block. Note that although this is not represented in the simulator, in this figure every block's "dangerousness" is colour-coded.

Simulation Optimizations

While these are not one of the core aspects of our simulator, various code optimizations were implemented throughout the development course of the simulator. In fact, without them it would have been impossible to achieve the referred running times, so we believe it is

important to mention the most relevant optimizations so as to contextualise this effort and not lead future readers into the misconception that implementing such a system is merely stripping down the native code to its data structures and logic.

Partial game world representation, event trigger abstraction, AI minification and in-memory player models as simulation optimizations

One of the earliest optimizations was implementing, as in Vanish, only a partial representation of the game world. This limited future block generation conflicts and reduced memory and CPU usage. In a second phase, this optimization was improved by segregating the various components of the gamestate into the aforementioned layered representation.

Another early optimization was to abstract all of the asset anchoring and visibility check code, simplifying the latter to a grid-based radius check. The same was done in regards to sound propagation and events' variable trigger timing intervals. All of these operations are considerably costly in a 3D environment. However, since we were transferring the game world into a 2D, grid-based symbolic representation, they were no longer necessary and saved us precious CPU time.

While we could have also ported the game's native AI code, this would have been inefficient as many elements of the creature's behaviour, such as path-finding and navigation[21] had become unnecessary. Creature encounters were abstracted to a sort of block-based game event which the player had a specific probability of escaping unharmed depending on the creature's AI state (see next sub-section).

Given the large configuration space allowed by Vanish's procedural engine, it was also necessary to define how it should be sampled; too spaced apart and we missed potentially interesting configurations, too tightly and the simulation time would exceed any reasonable time budget. This trade-off itself is not an optimization but the process that led us to define the sampling rates for each dimension of the configuration vector most certainly is.

Finally, perhaps the most relevant optimization was done in the later stages of the simulator's development and concerned the affective reaction models. Consulting the models using their formal descriptions implied solving a series of polynomial equations for every player cluster which, over more than a half a million simulation cycles added up to a substantial percentage of the simulation's total CPU time. The solution to this was to map all of the models offline and inject said map into memory during the simulator's bootstrap process. This made estimating

[21] Not necessarily the same as the former deals with finding the path and the latter also with how to traverse the path, animation-wise included.

player's reactions $O(1)$, drastically reducing the simulation runtimes by almost ~40%.

Simulating Players' Gameplay Styles

Modelling players' gameplay style through probabilistic reasoning based on empirical data

While we have already addressed the (arguably) more complex problem of how to predict players' emotional reactions to game events, some aspects of their gameplay experience remain unmodelled; namely their actual gameplay style. This entails how they navigate the game world and act when (inter)action is required – e.g. do they run from the creature as soon as they see it, do they explore visually different or more well lit areas of the world, etc. In other terms, we're interested in modelling their decision process when faced with more than one alternative. More precisely, given the nature of our simulated world, we'd like to understand how players *1)* decide which direction to take when faced with a crossroads, *2)* react when faced with the creature, and *3)* make use of the evasion tunnels (both when exploring and when being chased by the creature). Although we could model many other aspects of how players move about the game world (e.g. when they use the light sticks or how often they run), this is not all too relevant for our simulation purposes as it has relatively low impact on the game's outcome. It would make sense were we interested in recreating a visually accurate and believable representation of player's behaviour. However, that would be a completely different – much more challenging – endeavour that is completely outside the scope of this experiment or thesis.

To model the aforementioned aspects of player behaviour we analysed players' through direct observation during both the Vanish studies and on the various online gameplay footage available on Youtube (see footnote 11 in Chapter IV). We also interviewed some of our study participants to understand why they adopted a specific gameplay style or strategy.

A simple maze exploration strategy turned instinctual and reactive due to Vanish's procedural game world generation

Navigationally wise, most players approached the game as if they were solving a maze; always turn in the same direction. This is a popular solution to solving mazes as the player is bound eventually find the exit if he follows a common wall. However, as players quickly understood the game was procedurally generated and the maze's layout changed, they abandoned this strategy, opting to choose their destination based on environment events (i.e. avoiding dimly lit areas or creature sound effect sources). Given this feedback, we modelled this decision process using a few basic rules that stated that the player AI should avoid low illumination areas and creature sound sources but

should be attracted towards other game world environment assets (e.g. notes posted in walls or the machinery sound effects of objective rooms). In the absence of these influencing factors, the choice was made pseudo-randomly with an (empirically obtained) 40-25-30-5 bias towards left, front, right and backtrack options. Do note that while the player AI has free choice over where to go, its importance is of relatively low value as we are able to easily change the game world's layout. It does, however, add some realism to the simulation as in some cases we might not be able to generate specific game events or world blocks depending on how the player AI chose to move to.

Dealing with creature encounters based on context and historic data

Regarding the second aspect of the player AI's gameplay style – how it deals with creature encounters – we also took a probabilistic approach based on empirical data. When faced with the creature, since the player has no means to fight back, he two options: stay still in hopes the creature can't locate him or run from it. Based on the gameplay data we collected, we estimated that players had roughly a 75% chance of surviving their creature encounters. However, this number was considerably influenced by the creature's AI state (passive, aggressive, etc.) and by the players' positioning in the game world relative to the creature (i.e. at a crossroads versus in a dead-end). To more accurately simulate these encounters, the aforementioned base survival probability was coupled with a modifier value that took on the following values:

- Player positioning: If the player faces the creature head-on, the probability remains unaltered. If the creature blind sights the player in a corner or comes up behind him, it is reduced by 30%. If the player is at a dead-end, there is a coin flip to determine whether the player moves or stays still. If he moves, he dies. If he stays still, it depends on the creature's AI state; passive, he lives, otherwise, he dies.
- Creature AI: If the player is not at a dead end, then the encounter's outcome is further influenced by the creature's AI too. If it's passive, the survival probability remains unchanged, otherwise it drops 40% and a "chase mode" ensues for the following 3 world blocks. During the "chase mode", the player must run or reach an evasion tunnel to avoid getting caught. During this "chase mode" fails if the player has a sanity level below 10%, or if he runs out of stamina (i.e. becomes unable to run).

Finally, we also chose to model how often players made use of evasion tunnels as they appeared in the game world since, despite players not being aware of this, whenever an evasion tunnel is used, the game world is reset. This applies to game blocks and events, as well as

creature spawn events and stalking modes, effectively making the game easier.

Simulation Modes

As we have previously discussed, the simulator was built with two experiments in mind. The first one, meant to test whether we could automatically infer which game configurations were more suited to elicit a specific emotional state in an offline manner – henceforth referred to as Offline Biofeedback Optimization (OBO). And a second one to perform adaptations during actual gameplay – henceforth referred to as Online Regulated Biofeedback (ORB).

Simulating modes for both offline and online game experience optimization

OBO works by taking the game configuration parameters in the simulator's configuration file and generating all possible configuration vectors. It then runs each configuration vector the specified number of repetitions for each player, flushing the simulation's output periodically to the log file. As it is merely using the players' affective reaction models to assess their reactions to game events, it does not care about whether the simulation is progressing towards any specific target emotional state, leaving that to *a posteriori* analysis.

ORB, on the other hand, needs to plan ahead towards a specific target emotional state or pattern. This implies that: *1)* it makes a much heavier use of the players' reaction models and executes additional simulation steps, making the computational cost much higher than OBO, and *2)* since it needs to adapt the simulation towards a specific emotional state or pattern, it must repeat the simulation for each target affective experience.

Despite both being conceptually very similar and making use of the same simulation elements, this meant that ultimately, we had to confine the ORB experiment to a limited number of game configuration vectors. Our experimental protocol was thus to use the OBO experiment to identify the best game configuration vectors for each target emotional state and pattern (see next section) and then test whether using ORB's more intrusive gameplay adaptation mechanism, we were able improve on these results.

7.2 EXPERIMENTAL PROTOCOL

As mentioned in the previous section, given the intractability of running both the OBO and ORB experiments for all game configuration vectors and target emotional states, we decided to use the OBO experiment to

identify which configuration vectors were optimal for each target state. We then used both the optimal configuration vector and the game's original configuration vector in the ORB experiment to assess whether we were able to generate any improvements over OBO's offline optimization. In this section we will describe the results obtained for both experiments, focusing on the differences between both gameplay optimization methods in respect to each other and the non-biofeedback regulated version of the game. We also present a brief analysis of the effects of altering the gameplay adaptation frequency on the ORB experiment.

Prior to presenting our results we must still answer two relevant questions: *1)* "*What emotional states should we elicit?*", and *2)* "How *to best measure the 'goodness' of a simulation over time?*".

Target Emotional States

What types of affective experiences should we target and how to map them to the AV space?

The answer to our first question is that being a survival horror game we would wish to elicit rather negatively valenced experience. However, several other emotional states are present when playing a game and have been previously tied to a good gameplay experience (see Chapter II). Thus, we chose to elicit a set of 16 emotions we felt accurately described the most relevant states of mind when playing a game. Since Russell's AV space does not classify emotional states as specific emotions we turned to (Hepach, Kliemann, Grüneisen, Heekeren, & Dziobek, 2011) to define a mapping between emotional states and their corresponding emotion. Gameplay is, however, characterised for being a non-static experience. Measuring our gameplay adaptations solely based on our ability to elicit specific emotional states, disregarding how well we are able to force a transition between them, would be insufficient. Thus, we decided to also measure our ability to elicit emotional patterns (see Figure 7:1). The tested patterns were a rather simple sinusoidal wave for both arousal and valence on the higher and lower scale values, respectively. Three different patterns were created using different wavelengths of 2, 3 and 5 minutes. For a list of the considered emotions / emotional patterns see Table 7:1.

Defining Fitness

To answer our second question we used a traditional distance metric between the target emotion / emotional pattern and the players' emotional state over each simulation. This distance metric provides us with a proxy measurement to the fitness of each configuration vector in respect to the target emotion / emotional pattern.

How to measure the "goodness" of the obtained results?

Due to the wide emotional range present in the chosen emotions and emotional patterns, as well as the comparably lower fluctuation of the emotional states[22], a linear function was likely not the most appropriate mapping function. Moreover, since the main purpose is to highly differentiate the good from the bad emotional experiences, a sigmoid function was better suited to more accurately differentiate between closely positioned emotional states. A brief formalization follows.

And how to do so such that increasingly modest improvements are still noticeable?

Let p be a point in a two-dimensional Euclidean space, such that $p=(p_1,p_2)$, where $p_1, p_2 \in \mathbb{R}$ represent p's first and second dimensional coordinates. Furthermore, let q^2 be the quadratic distance between two coordinates c and c', such that:

$$q^2_{c,c'} = (c - c')^2 \qquad \text{(Eq. 8:1)}$$

Furthermore, let d, the weighted Euclidean distance between two points p and p', be given by:

$$d_{p,p'} = \sqrt{\alpha\left(q^2_{p_1,p'_1}\right) + \beta\left(q^2_{p_2,p'_2}\right)} \qquad \text{(Eq. 8:2)}$$

Where α and β are weighting parameters for each of the Euclidean dimensions, such that $\alpha,\beta \in \mathbb{R} \wedge \alpha,\beta \geq 0 \wedge \alpha + \beta = 2$. These weighting parameters are meant to favour or penalize each emotional dimension, according to the perceived difficulty in adjusting it. For example, valence presented itself notoriously more difficult to elicit in our previous study in Chapter IV. Hence, a higher weight might be attributed to it. Since we wish to perform an unbiased analysis, in this chapter, $\alpha = \beta = 1$.

The fitness of a certain point p_c to a target point p_t is given by f and is inversely correlated to its distance from p_t. Since we aimed at penalising values further from p_t, it can be trivially concluded that a linear correlation function between distance and fitness would not be adequate. Thus, f is given by a sigmoid function of the form $f(x) = {x}/{\sqrt{\zeta + x^2}}$:

$$f_{p_c,p_t} = \begin{cases} 1 & ,if\ d_{p_c,p_t} = 0 \\ 1 - \dfrac{d_{p_c,p_t}}{\sqrt{\zeta + {d_{p_c,p_t}}^2}}, & otherwise \end{cases} \qquad \text{(Eq. 8:3)}$$

With $\zeta \in \mathbb{R}$ being its exponential tuning parameter – in this case empirical analysis revealed $\zeta = 200$ to be a suitable value. It should also

[22] Which was to be expected given the nature of the game and our prior results in Chapter IV.

be noted that the fitness value f is *a posteriori* normalised in the range [0,1] using the minimum fitness value possible in the AV space (i.e. $d \cong 14.14$).

7.3 EXPERIMENTAL RESULTS

As previously discussed, as the simulations are run, the players' emotional states over each experiment are logged to an external text file. Out of these values, what we are most interested are players' arousal and valence levels over time, as well as their fitness values in respect to each selected emotion and emotional pattern. Using this data, we are able to rank each configuration vector according to their fitness values (OBO experiment) and assess the emotional elicitation capabilities of our affective reaction models (ORB experiment). In this section we present these results in the form of several plots and tables depicting the general performance obtained in each experiment. We begin by analysing OBO's results and then move on to compare its results to ORB.

Offline Biofeedback Optimization (OBO) Results

Regarding the OBO experiment, we present a study of its fitness values across the selected game configuration subspace as it provides a high-level but complete overview of its improvements over the game's vanilla configuration.

Optimizing the game's configuration vector towards specific affective experiences via offline configuration space searches

Due to the large size of the configuration subspace used in our simulations and since our goal is to perceive discernible differences in fitness, we chose to present a representative sample of the (sorted) fitness values obtained by the configuration vectors. The presented configuration vectors are thus sorted according to their mean fitness value and sampled using the fitness distribution quintiles. The vanilla configuration vector is also shown as a baseline comparison. As mentioned at the beginning of this section, the "2 Minutes", "3 Minutes" and "5 Minutes" keywords represent the considered emotional patterns; 2, 3 and 5 being the pattern's period.

Table 7:1. Fitness improvements along sorted game configuration vectors for the OBO experiment.

	Best	25%	50%	75%	Worst	*Vanilla*
Anxious	0,6548	0,6437	0,6405	0,6374	0,6270	0,6412
Bored	0,6004	0,5932	0,5888	0,5815	0,5592	0,5899
Concerned	0,9394	0,9353	0,9336	0,9313	0,9261	0,9334
Confident	0,3019	0,2976	0,2960	0,2934	0,2865	0,2950
Confused	0,8537	0,8456	0,8424	0,8380	0,8269	0,8447
Desperate	0,4739	0,4656	0,4631	0,4606	0,4529	0,4640
Enthusiastic	0,4435	0,4354	0,4323	0,4297	0,4220	0,4300
Frantic	0,4045	0,3955	0,3918	0,3894	0,3830	0,3929
Frustrated	0,5295	0,5224	0,5202	0,5172	0,5086	0,5215
Jumpy	0,6562	0,6444	0,6404	0,6373	0,6278	0,6406
Proud	0,3985	0,3910	0,3883	0,3849	0,3744	0,3866
Shocked	0,5455	0,5349	0,5311	0,5282	0,5197	0,5318
Surprised	0,7112	0,6955	0,6901	0,6842	0,6666	0,6857
Triumphant	0,5730	0,5606	0,5562	0,5516	0,5378	0,5529
2 Minutes	0,6957	0,6899	0,6881	0,6860	0,6791	0,6881
3 Minutes	0,7063	0,7009	0,6991	0,6970	0,6903	0,6990
5 Minutes	0,6801	0,6725	0,6698	0,6675	0,6600	0,6700

As can be seen in Table 7:1, the game's vanilla configuration vector presents similar performance to that of the mean configuration vector. This was to be expected as the configuration subspace was obtained by deviating from the game's original configuration and seems to imply our simulations are indeed stable enough to be considered accurate. Moreover, the range of fitness values for each of these emotions / emotional patterns varies more widely when the fitness value is neither in the lower or upper ranges of the spectrum. Emotions such as "Enthusiastic" and "Shocked" reveal a larger range of fitness values than, for example, the "Concerned" emotion. This appears to be a direct by-product of the fitness function being drawn from a sigmoid type, thus attributing a larger boost in fitness as values become closer and closer to the desired emotional state.

Regarding (static) emotions, the foremost observation is that we are not able to significantly optimize the gameplay experience towards some emotions not easily found on survival horror games. Some examples include the "Confident", "Proud" and "Triumphant" emotions. As these emotions are not common, or make sense other than briefly, in this game genre, these low fitness values were to be expected and, in our opinion, constitute validation that our simulator is not artificially producing unrealistic results. On the other hand, emotions such as "Concerned", "Confused" and "Surprised", which are common in the emotional atmosphere generated by horror games, present a high degree of fitness values.

Emotional patterns are an elusive target for static, offline gameplay optimization

A more in-depth analysis allows us to extend this line of thought to the AV space as a whole. Emotions or emotional states located near in the second quadrant of the AV space (high arousal, low valence), present significantly higher fitness values. Fundamentally, this relates to the nature of Vanish's game genre and the emotional reactions it is able to elicit. Being a horror game, it is considerably harder to create emotional responses that elicit "happy" (high valence) or "relaxing" (low arousal) emotions. In fact, a "happy" or "relaxing" horror game is a nonsensical proposition. Even if that was our goal, in such a situation, the most that we could do would be to not stimulate the player and allow him to decay into his base emotional state. However, doing that would most likely lead to a very boring gameplay experience.

Similarly, emotions outside of the game's intrinsic emotional spectrum are not easily elicited

A more subtle implication of this inability to elicit high valence or low arousal states is that, unless we allow for controlled decay periods, no game configuration vector can, consistently over various simulations, display a dynamic affective experience (i.e. create an emotional pattern). Since given Vanish's probabilistic procedural generation mechanics and lack of players' individual responses, controlled decay periods are not a possibility. Thus, an offline simulation is, sadly, unable to successfully generate emotional patterns of any kind.

Online Regulated Biofeedback (ORB) Results

Perhaps the most simple and effective way to compare the OBO and ORB experiments is to visualize the fitness of both throughout the same emotion optimization simulations, as shown below in Figure 7:3.

Addressing OBO's limitations through online gameplay experience adaptation mechanisms

Understanding the effects of the ORB experiment's dynamic emotional regulation mechanism however, requires that we examine the evolution of the fitness values over the course of the simulation for each emotion / emotional pattern. The plots presented in Figure 7:4 illustrate the average fitness values of all modelled study participants for each considered target affective experience. Each plot portrays the game's

vanilla configuration vector and the best configuration vectors for each experiment. This allows us to compare the fitness improvement of each experiment over the game's original version. In addition, both experiences can also be compared in terms of their presented fitness. As can be seen, the ORB experiment displays overall better results, which was to be expected given its augmented capabilities in adapting the gameplay experience based on players' affective reaction models. In a similar note, the difference between the game's vanilla configuration vector and the best configuration vectors (selected individually per player, not the overall population) is evident on both experiments as a notorious increase in fitness is visible.

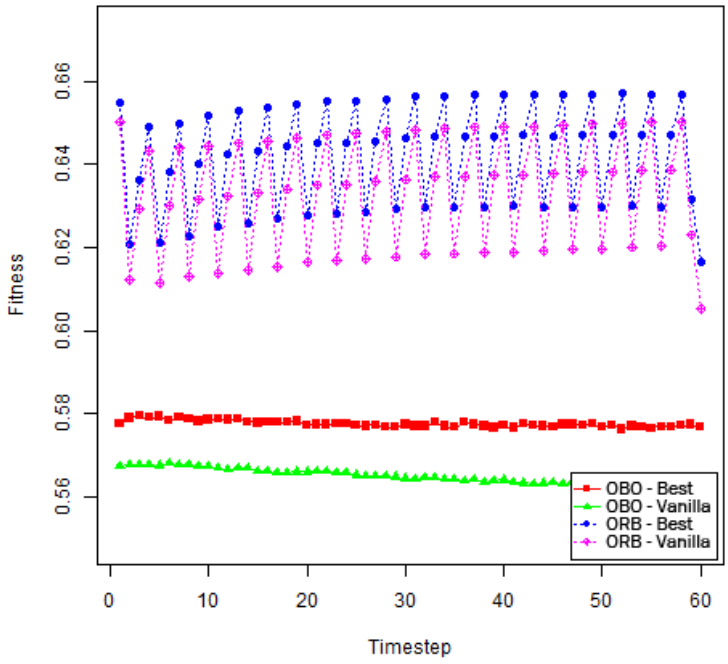


Figure 7:3. Fitness comparison between OBO and ORB experiments for the "Jumpy" emotion.

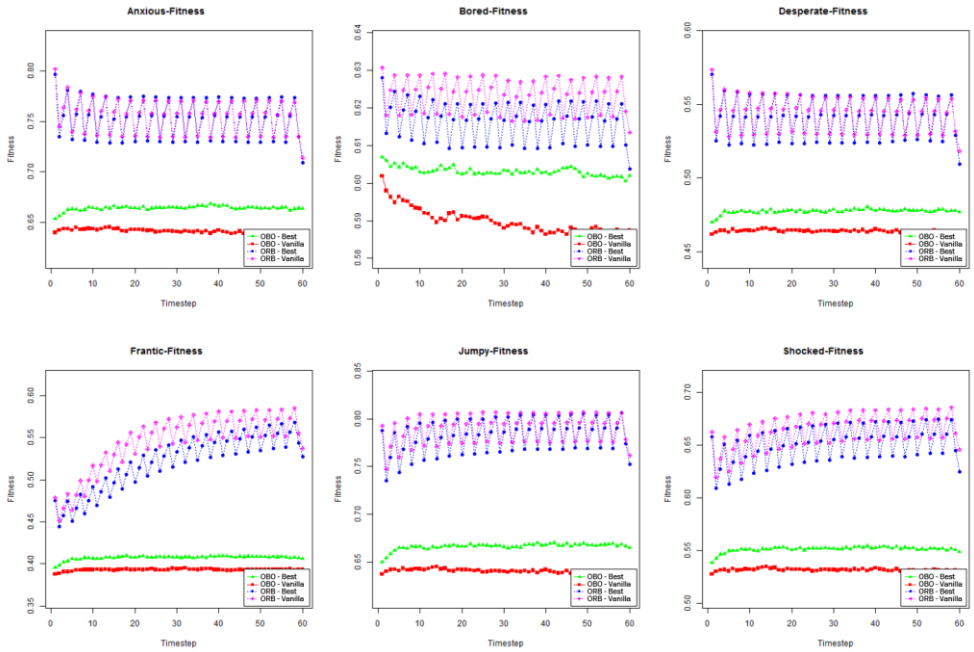


Figure 7:4. Fitness comparison between OBO and ORB experiments for the emotions relevant to the horror game genre.

Notice the "jaggy" behaviour on the ORB experiment. This is not an experimental error. As stated in the previous sections, the ORB simulation mode works by selecting a series of game events to stimulate the player in the direction of the desired emotional state at specific intervals. Given that these intervals occur at the same simulation timesteps, it is only natural that on those timesteps there is an increase in fitness due to the introduced changes in gameplay. This fitness increase then subdues until the following adaptation timestep due to the emotional decay mechanism. This is not to say that these "fitness spikes" do not occur in the OBO experiment (or in the game's actual gameplay). As we have seen in previous chapters, these drastic changes in pacing are what make the game interesting and challenging and the emotional reactions they create are our model's main data source. However, in the OBO experiment, these changes do not occur at fixed time intervals (at least not in a fixed schedule as in ORB) and over the course of hundreds of thousands of simulation timesteps, they eventually average out to nearly the same values, as can be seen in Figure 7:4. This effect is naturally also present in both the arousal and valence dimensions, as it is where it stems from (see Figure 7:5). At its core, this "jaggy" behaviour is an experimental artefact originating from the adaptation intervals. Smoothing the simulation's output would have eliminated it but as it stands, these "spikes" are a (visual) testament to ORB's emotional elicitation capabilities, which further strengthen our

"Affective jaggedness" is as much of a characteristic trait of game playout adaptation as it is a hallmark of good gameplay design

thesis that emotionally regulated biofeedback experiences are in fact capable of significant emotional elicitation.

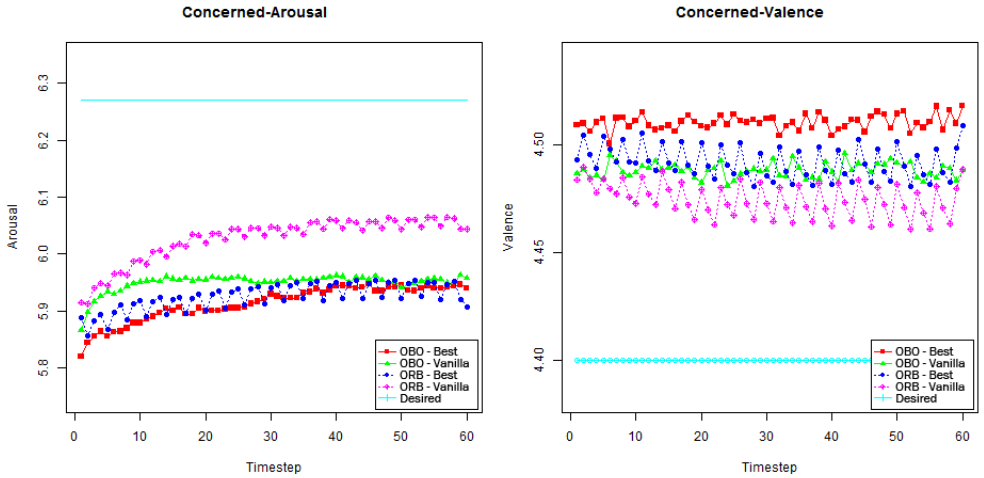


Figure 7:5. Jaggy behaviour observed in the arousal and valence dimensions due to ORB's adaptation mechanism.

Regarding a more objective analysis of the two experiments, we consider it best to compare the fitness values achieved by them across each target emotion and emotional pattern. We present this comparison bellow, in Table 7:2.

An objective comparison of the fitness levels achieved by both optimization experiments

Table 7:2. Fitness comparison between the OBO and ORB experiments for the selected target affective experiences using the vanilla, overall (population) and individually best configuration vectors.

		Vanilla		Best Population		Best Individual	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Anxious	ORB	0,7534	0,0829	0,7597	0,0762	0,7558	0,0764
	OBO	0,6412	0,0473	0,6548	0,0551	0,6639	0,0576
Bored	ORB	0,6165	0,0336	0,6237	0,0282	0,6235	0,0281
	OBO	0,5899	0,0318	0,6004	0,0265	0,6031	0,0270
Concerned	ORB	0,9462	0,0409	0,9511	0,0367	0,9412	0,0429
	OBO	0,9334	0,0418	0,9394	0,0375	0,9536	0,0399
Confident	ORB	0,3163	0,0234	0,3229	0,0229	0,3227	0,0227
	OBO	0,2950	0,0141	0,3019	0,0148	0,3023	0,0149
Confused	ORB	0,8853	0,0770	0,8926	0,0739	0,8847	0,0774
	OBO	0,8447	0,0574	0,8537	0,0573	0,8667	0,0584

Desperate	ORB	0,5413	0,0579	0,5457	0,0520	0,5441	0,0528
	OBO	0,4640	0,0311	0,4739	0,0358	0,4774	0,0379
Enthusiastic	ORB	0,4707	0,0467	0,4800	0,0444	0,4797	0,0447
	OBO	0,4300	0,0239	0,4435	0,0277	0,4447	0,0267
Frantic	ORB	0,5208	0,0647	0,5442	0,0702	0,5441	0,0699
	OBO	0,3929	0,0262	0,4045	0,0323	0,4072	0,0337
Frustrated	ORB	0,5931	0,0715	0,5929	0,0658	0,5884	0,0666
	OBO	0,5215	0,0357	0,5295	0,0398	0,5349	0,0403
Jumpy	ORB	0,7820	0,0755	0,7927	0,0707	0,7900	0,0719
	OBO	0,6406	0,0468	0,6562	0,0563	0,6663	0,0602
Proud	ORB	0,4226	0,0407	0,4310	0,0383	0,4315	0,0380
	OBO	0,3866	0,0231	0,3985	0,0241	0,3994	0,0251
Shocked	ORB	0,6504	0,0668	0,6628	0,0622	0,6642	0,0638
	OBO	0,5318	0,0380	0,5455	0,0452	0,5513	0,0492
Surprised	ORB	0,7560	0,0816	0,7703	0,0723	0,7699	0,0735
	OBO	0,6857	0,0459	0,7112	0,0455	0,7132	0,0490
Triumphant	ORB	0,6118	0,0677	0,6233	0,0610	0,6234	0,0619
	OBO	0,5529	0,0365	0,5730	0,0370	0,5746	0,0397
2min	ORB	0,7524	0,1733	0,7580	0,1602	0,7554	0,1550
	OBO	0,6881	0,2122	0,6957	0,2057	0,7022	0,2029
3min	ORB	0,7724	0,1686	0,7798	0,1554	0,7770	0,1503
	OBO	0,6990	0,2120	0,7063	0,2059	0,7141	0,2037
5min	ORB	0,7797	0,1534	0,7958	0,1397	0,7942	0,1357
	OBO	0,6700	0,2010	0,6801	0,1962	0,6875	0,1937

Two important deductions can be taken from Table 7:2. First, that the ORB experiment presents significantly better fitness values across all target affective experiences. This was to be expected given our previous analyses but here lies objective proof of it. Secondly, that it achieves this improved fitness in a more consistent, stable manner, as can be seen by the presented lower standard deviations. This means that not only is the adaptation mechanism present in ORB able to achieve affective experiences closer to the ones that were idealised, it is also able to do it for a larger number of players and/or in a more

controllable fashion. Also note that, as in the OBO experiment, the ranges in which the fitness values are presented vary from target affective experience.

To allow for a more manageable comparison between both experiences, a global overview was computed based on the average fitness values over the target affective experiences. However, since there is an intrinsically different nature between the static (emotions) and dynamic (emotional patterns) target affective experiences, this distinction was maintained. This global overview is presented in Table 7:3.

Overall, not only is ORB able to produce higher fitness values, it is able to do so more consistently

Table 7:3. Fitness comparison between the OBO and ORB experiments for the vanilla, overall (population) and individually best game configuration vectors.

		Vanilla		Best Population		Best Individual	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Emotions (Static)	ORB	0,6333	0,0594	0,6423	0,0553	0,6402	0,0565
	OBO	0,5650	0,0357	0,5776	0,0382	0,5828	0,0400
Emotional Patterns (Dynamic)	ORB	0,7682	0,1651	0,7779	0,1518	0,7755	0,1470
	OBO	0,6857	0,2084	0,6940	0,2026	0,7013	0,2001

This improvement is, however, much more significant in regards to the elicitation of emotional patterns

Table 7:3 allows us to, again, draw several conclusions. Firstly, that the increase in fitness is larger for dynamic affective experiences (emotional patterns). Given our conclusions on OBO's limitations in this regard, this comes with little surprises. It also appears that these present a bigger challenge, even for ORB, as the presented standard deviation is an order of magnitude larger. Since we are not optimising for a particular emotional state and there is also a timing effect due to the target emotional state changing over time, we believe it is safe to say that this is a very successful outcome for ORB. Similarly to what we have previously mentioned in OBO's discussion, some emotions presented low fitness levels because the target emotional states were simply outside of Vanish's emotional spectrum. Presumably, this would still be a limitation in ORB's case since a combination of game events that could elicit these emotional states would continue to not exist. A closer examination of Table 7:2 confirms that despite its ability to ameliorate the situation, this assumption holds true for ORB.

Overall, it can be noted that the ORB experiment presented overall higher fitness and stability values, independently of the target affective experience. This implies that, not only did ORB present better results, it produced "tighter", less erratic affective experiences over time. This

increase in stability is even more important for dynamic affective experiences, as a relatively small deviation in the emotional pattern's overall waveform might have a, presumably, bigger impact.

A brief analysis on the discrepant difficulty in arousal and valence elicitation

Since fitness is still an abstract concept that does not properly illustrate the individual differences in the considered emotional dimensions, its analysis is not sufficient to fully understand the obtained results. Hence, Figures 7:6 and 7:7 showcase the observed distance between the obtained and desired arousal and valence levels for each target affective experience on both experiments.

Given the differences between the simulated and desired emotional states portrayed in Figures 7:6 and 7:7, it again seems evident that the ORB experiment produces affective experiences closer to the desired ones. Furthermore, on closer inspection, it appears arousal presents larger variances between the two experiments, while the valence dimension expresses almost identical results between both experiences. This curious occurrence is justified due to the difficulty in eliciting significant valence responses. This can be attributed to the game's rather lacking hedonic expressivity (i.e. game events are predominantly negative, with the only calming aspect of the experience being the evasion tunnels and, arguably, the absence thereof said events). Arousal, on the other hand, appears not to suffer from this ailment. An interesting observation is that arousal displays similar values when the desired emotions have low target arousal. Given the game's previously discussed (virtual) inability of calming the player other than by limiting the number of game events, it would appear these values are caused by hitting the minimal arousal allowed by the game's emotional spectrum.

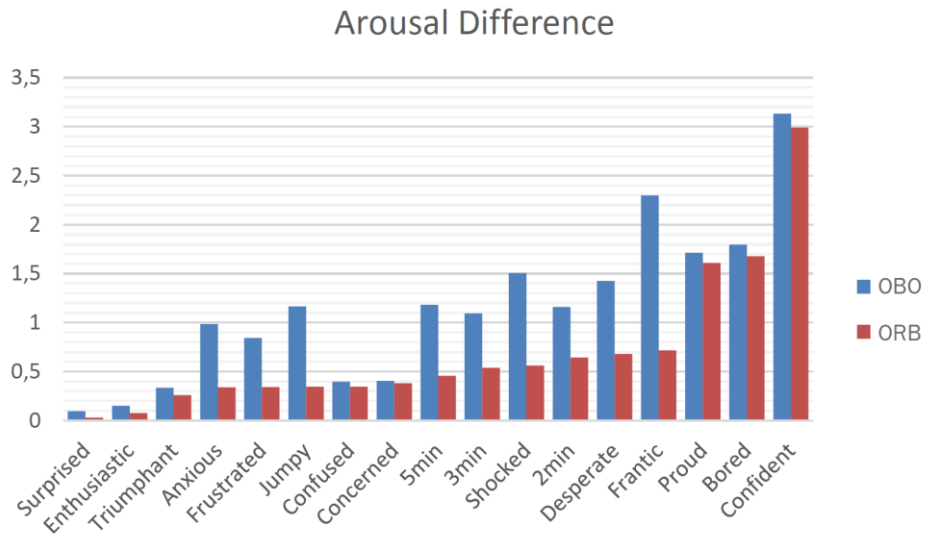


Figure 7:6. Arousal distance from target emotional states on OBO and ORB experiments.

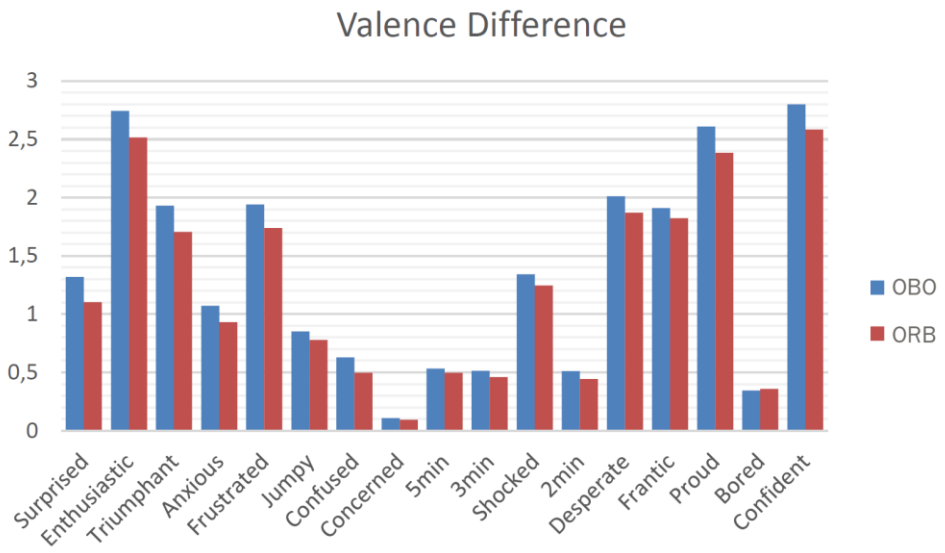


Figure 7:7. Valence distance from target emotional states on OBO and ORB experiments.

On a final effort to thoroughly examine the presented results, we created several plots that show the mean AV values in comparison with the target affective experience. We expect these provide a degree of visual aid and properly illustrate and visually summarise the capabilities of this technology for both offline and online applications.

A high-level view of fitness values on both experiments over the simulations' course

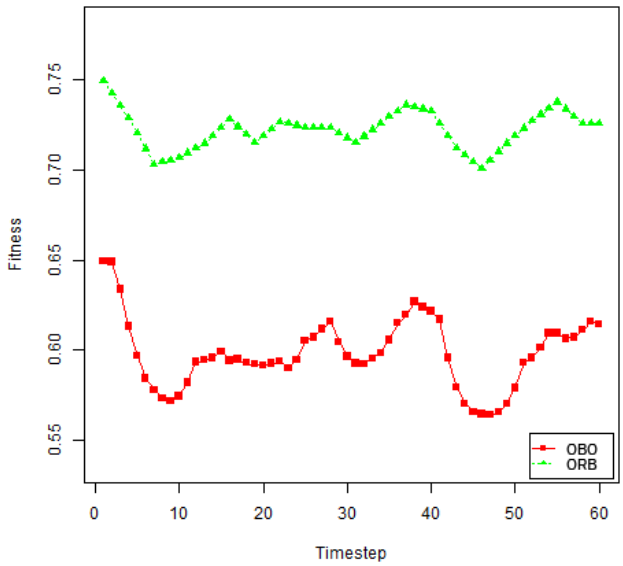
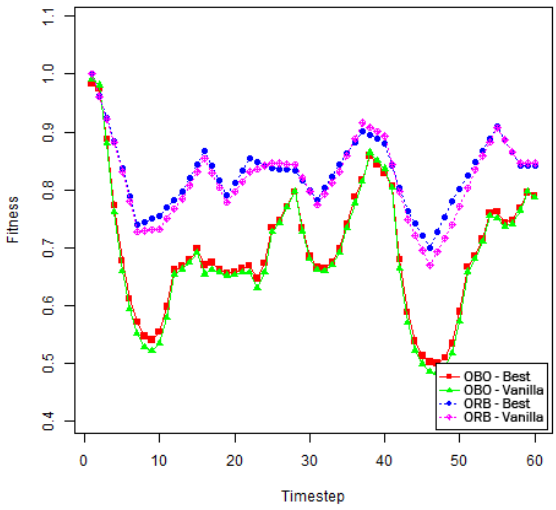


Figure 7:8. Mean fitness over simulation timesteps for the OBO and ORB experiments.

Figure 7:8 shows the mean fitness over simulation timesteps for both the OBO and ORB experiments. A clear distinction exists between both the absolute fitness values offered by each experiment. Moreover, the results presented by the ORB experiment are also, as previously discussed, noticeably more stable.



A closer look the differences between offline and online optimizations towards emotional patterns and how fitness alone fails to capture their dynamic nature

Figure 7:9. Mean fitness over simulation timesteps for the OBO and ORB experiments for targeted dynamic affective experiences.

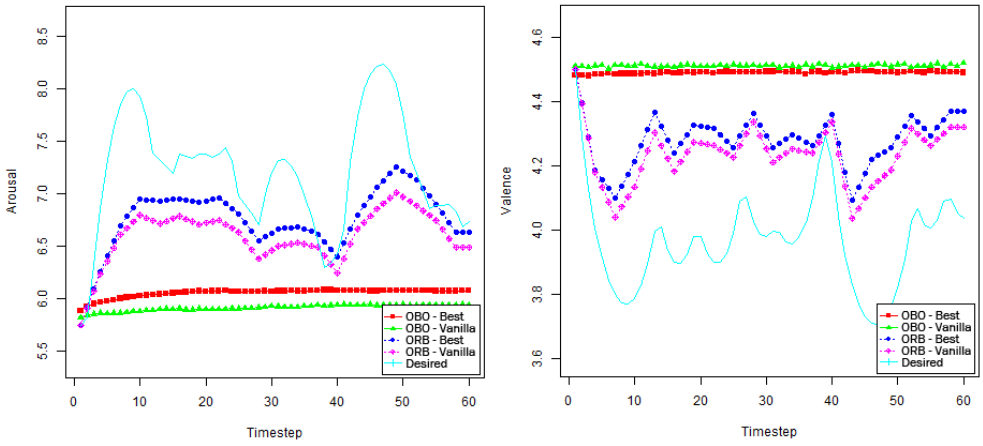


Figure 7:10. Mean arousal (left) and valence (right) values for the dynamic emotional states for both the OBO and ORB experiments.

On a similar note, the plot shown in Figure 7:9 presents the mean fitness values for all target emotional patterns combined (hence the characteristic waveform). While ORB's better performance is noticeable in Figure 7:9, its adaptive behaviour is not. However, if we consider this same plot for the AV dimensions instead of overall fitness, it becomes wildly apparent, as we can see in Figure 7:10. This adaptive behaviour is what distinguishes ORB from OBO, giving it the edge in not only

eliciting emotional patterns but, ultimately, in its ability to adapt to player's ever changing reactions to the surrounding medium.

The effects and importance of appropriate gameplay adaptation periods

Given the improvements observed in the ORB experiment, we became interested in how much adapting the intrusiveness value could influence the obtained results. Recall that the simulator's intrusiveness parameter is the period at which – in the ORB experiment – the gameplay experience is adapted according to the best set of game events found by the player's affective reaction model.

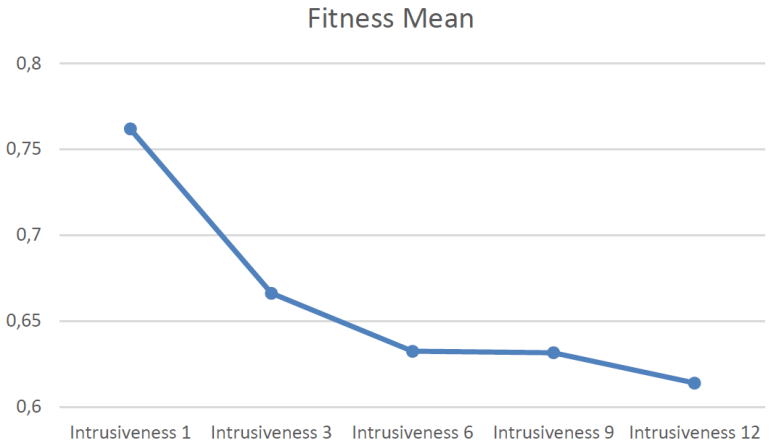


Figure 7:11. Mean fitness values for the ORB experiment over gameplay adaptation intrusiveness.

Despite the positive correlation between shorter adaptation periods and fitness values, the impact of too much adaptation can be even more devastating than too little

As expected, there seems to be a clear correlation between the intrusiveness parameter and overall fitness. This indicates that not only does dynamically generating game events produces better results than random generation but that setting the adaptation interval too high will cause virtually no effect (as can be seen from the fitness values for intrusiveness 6, 9 and 12). A quick analysis of Figure 7:11 reveals an inverse exponential correlation function which would seem more sensible than a direct correlation as with increasingly higher intrusiveness intervals, the benefits quickly approximate to zero. Despite this encouraging result, it would be advisable to exercise caution on its interpretation as using a maximal intrusiveness (adjusting the game every 10 seconds or even less) can quickly make the gameplay too erratic or hectic with no conducting line. While this might work in some games, in the limit it almost always configures bad gameplay design as there is a need for some pacing for players to enjoy the game and sense they are making progress. In our particular case, this could also result in overfitting, thus biasing our results. As such, this analysis, while presented as a final consideration on this section, was a major factor in choosing an intrusiveness value of 3 (once every 30 seconds). In a real-world application we would advise a mix of

intrusiveness levels 3 and 6, each with its own purpose; the former for more phasic game events (e.g. generating enemy placements and game assets), and the latter for tonic aspects of the gameplay experience (e.g. level layout, atmosphere and overall pacing).

7.4 DISCUSSION

The results presented in this chapter regarding the affective reaction model's ability to elicit specific target emotional states are encouraging. While this technique is not a "silver bullet" for emotionally regulating affective experiences, it does significantly improve on past results, such as the ones presented in (S. Gilroy & Porteous, 2012; Holmgård et al., 2013; Martinez et al., 2011; Moreira, 2010; Shaker et al., 2010; Tognetti et al., 2010; Wang & Marsella, 2006). Besides indicating that we were able to successfully adapt the gameplay experience towards specific emotional states, it also demonstrates that we were able to do so in a stable fashion and for dynamic emotional states.

Emotional regulation is not a silver bullet

In our first experiment, we saw that it was possible to optimise the game's configuration parameters towards a specific (static) emotional state with consistently high fitness values over our study population. However, the drawback to this approach was that despite being very CPU intensive (a potential issue when scaling this approach to more complex games), it was unable to elicit emotional patterns. It is possible that using dynamic gameplay adaptation mechanisms such as the ones in Chapter IV would ameliorate this issue on a practical implementation. However, it would still require a real-time feed on the players' physiological/emotional state. In such a scenario, it would be best to simply use a more sophisticated approach that relied on player models (ORB) as all pre-requisites would already be satisfied. Having said this, the static adaptation mechanisms do have a place in acting as a stand-in for the model driven adaptation as in the early phases of the game's release no player data will be available or in academic research studies such as the ones presented by (Kivikangas et al., 2011; L. E. Nacke et al., 2010; L. E. Nacke, 2013).

Both offline and online gameplay optimization techniques have their place

As do static adaptation mechanisms as the ones in Chapter IV

In comparison with our first approach, by using players' models to guide the gameplay experience in real (simulation) time, we were subsequently able to improve on its limitations at various levels. Firstly, we were able to improve the achieved fitness and stability values for all static emotional states. Secondly, and most importantly, we gained the ability to elicit dynamic affective experiences. The added stability brought by the ORB approach also contributes to the method's applicability under tightly regulated environments. For example,

making sure it's output is optimally tailored and predictable during a phobia treatment session.

Despite ORB's ability to elicit emotional patterns, some emotional states remain out of reach

For all the added benefits of the ORB experiment, we were still faced with some limitations. Firstly and foremost, we found that despite its emotional elicitation capabilities, some emotional states remained outside of our reach as they are simply not within the game's emotional spectrum. This was to be expected and a good sign that our simulation is not producing artificially inflated or biased results. Secondly, similarly to what is discussed in Chapter IV, we confirmed that valence poses much of a bigger challenge in eliciting than arousal. In our opinion, this is mostly due to two factors: *1)* that the game is intrinsically biased towards low valence states, and *2)* that since valence is a more volatile and rapidly decaying state (perhaps derived from its more rational nature), it requires a more active adaptation process.

The fact that we chose to adopt a simulation approach was a difficult and rather dangerous decision. However, given the largely insurmountable logistic challenges, it was a necessary one. Despite this, we had no guarantee that we would be able to properly simulate the game's environment or players' emotional decay and gameplay style in a stable manner. Fortunately, given the achieved stable and in-line with previously observed physiological data, we have succeeded in generating reliable results and thus gained a very cost-effective and general purpose tool for future similar studies. Moreover, since we abstracted players' emotional states and game events, our simulation approach is not restrained to physiological data or gaming environments. As we have previously stated throughout this thesis, our choice of physiological input as our data source was merely due to its high precision. It is thus simple to envision how this approach could be used in conjunction with other, perhaps less intrusive data sources (e.g. facial recognition via a webcam stream, body movement, free text sentiment analysis, etc.), as long as these are interpreted in terms of a continuous theory of emotion.

Choosing a simulation based approach was risky but allowed us a much larger test coverage and created an additional contribution for the community

Most importantly, we have proved that this type of approach is feasible and is thus a viable alternative for commercial use. Having said that, we acknowledge that despite our encouraging results a new live study with a sample population orders of magnitude bigger is necessary to further validate these results. Another relevant aspect to be analysed is how effective these adaptation mechanisms are over time. While the affective reaction models are able to learn over time given new emotional reactions, we cannot be sure how these reactions will change over time. In other words, will the adaptation mechanisms be flexible

enough to overcome the game's loss in emotional elicitation capabilities over time?

7.5 SUMMARY

In this chapter, we have focused on two main points: *1)* whether the previously created affective reaction models could be used *offline* to optimize the game's configuration parameters towards a specific emotional state, and *2)* whether these same models could also be used to also elicit emotional patterns by doing real-time adaptations to the gameplay experience.

Chapter objectives and summary

Regarding our first question, through the OBO experiment we saw that it is indeed possible to optimise the game's configuration towards specific emotional states. However, this comes with a rather high computational price in sampling the whole game configuration space and, being a static optimization, is thus unable to elicit emotional patterns. This limitation was however overcome by the use of a dynamic adaptation mechanism in the ORB experiment, which boasted improved results across the board as well as increased stability.

Present contributions and achieved thesis objectives

Another key contribution present in this final chapter is our simulation platform, which was developed as a generic tool for games and affective computing research in general. With its development, we expect that in the future it can provide game designers and affective computing researchers with a common, stable platform for automating, reproducing and validating research results.

In terms of our thesis' objectives, in this chapter we have addressed objectives VI: *"Define an adaptation scheme to leverage the user collated data in order to reinforce a set of desired emotional states and patterns"* and VII: *"Test the created models / adaptation scheme to assess their general emotional elicitation capabilities on a wide range of emotional states"*. The formed refers, in general, to ORB's simulation flow and was addressed in the simulator's description at the beginning of this chapter. The latter essentially stated that we wished to test the effectiveness of the created models and, as such, is what the remainder of this chapter is dedicated to.

In sum, this chapter embodied the final validation study of our thesis and we expect these results will act as an original demonstration of the real-world applicability of physiological emotional regulation of human affective experience using human-computer interaction and artificial intelligence methodologies.

REFERENCES FOR CHAPTER VII

Gilroy, S. W., Cavazza, M., & Benayoun, M. (2009). Using affective trajectories to describe states of flow in interactive art. In Proceedings of the International Conference on Advances in Computer Enterntainment Technology - ACE '09 (p. 165). New York, New York, USA: ACM Press. doi:10.1145/1690388.1690416

Hepach, R., Kliemann, D., Grüneisen, S., Heekeren, H. R., & Dziobek, I. (2011). Conceptualizing emotions along the dimensions of valence, arousal, and communicative frequency - implications for social-cognitive tests and training tools. Frontiers in Psychology, 2(October), 266. doi:10.3389/fpsyg.2011.00266

Holmgård, C., Togelius, J., & Yannakakis, G. (2013). Decision Making Styles as Deviation from Rational Action A Super Mario Case Study. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment.

Matias Kivikangas, J., Nacke, L., & Ravaja, N. (2011). Developing a triangulation system for digital game events, observational video, and psychophysiological data to study emotional responses to a virtual character. Entertainment Computing, 2(1), 11–16. doi:10.1016/j.entcom.2011.03.006

Martinez, H. P., Garbarino, M., & Yannakakis, G. N. (2011). Generic Physiological Features as Predictors of Player Experience. In Proceedings of the 2011 Affective Computing and Intelligent Interaction Conference (pp. 267–276).

Moreira, V. H. V. G. (2010). BioStories Geração de Conteúdos Multimédia Dinâmicos Mediante Informação Biométrica da Audiência.

Nacke, L. E. (2013). An introduction to physiological player metrics for evaluating games. In Game Analytics (pp. 585–619). Springer London.

Nacke, L. E., Stellmach, S., & Lindley, C. a. (2010). Electroencephalographic Assessment of Player Experience: A Pilot Study in Affective Ludology. Simulation & Gaming. doi:10.1177/1046878110378140

Shaker, N., Yannakakis, G., & Togelius, J. (2010). Towards automatic personalized content generation for platform games. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (pp. 63–68).

Tognetti, S., Garbarino, M., Bonarini, A., & Matteucci, M. (2010). Modeling Enjoyment Preference from Physiological Responses in a Car Racing Game. In IEEE Conference on Computational Intelligence and Games (pp. 321–328).

Wang, N., & Marsella, S. (2006). Introducing EVG: An Emotion Evoking Game. In Intelligent Virtual Agents (pp. 282–291).

R E S E A R C H
S U M M A R Y

EMOTIONAL REGULATION OF INTERACTIVE EXPERIENCES

With the conclusion of the extensive simulation tests presented in Chapter VII, we have completed the final stage of our study on the usefulness of emotions towards regulating human affective experiences. As was our objective throughout this thesis, we have presented and verified the overarching hypothesis that:

> *"User experience in digital videogame environments can be influenced and enhanced by emotional regulation based on emotional state enforcing through a biofeedback loop. Furthermore, this loop can be implemented through the automatic evaluation of the relations between the user's emotional state data feed and an abstract stimuli set."*

Recalling the discussion present in Chapter I, given this hypothesis' complex nature, it was divided in four parts, each of which was tackled sequentially in Chapters III through VII:

I. Firstly, that even relatively simple, static gameplay adaptation mechanisms driven by players' emotional state result in (statistically) meaningful alterations in relevant user experience metrics (immersion, tension, flow, etc.).

II. Secondly, that by monitoring players' emotional states, it is possible to extract individual emotional reactions to game events in, at least, a semi-automatic fashion.

III. Thirdly, that these emotional reactions can be used to build affective reaction profiles through which players' reactions to future interactions can be estimated.

IV. Ultimately we hypothesize that using the aforementioned models, it is possible to build a system capable of regulating the user's emotional state via the presented emotional content.

Despite this effort to make our hypothesis more manageable by splitting it into these four sub-hypotheses, there were several obstacles that required our attention prior to any tests geared towards answering them. The prime examples of this were the development of a stable, validated, emotional detection system for use in real-time environments and a real-time adaptable case study (i.e. Vanish). As such, we delineated the following thesis objectives that define this thesis' addressed issues in a more fine-grained manner:

I. Propose a generic, conceptual architecture for the creation of emotionally adaptive systems (henceforth known as "Emotion Engine")

II. Define which existing state of the art methodologies are best suited for quantitatively and non-intrusively measure emotional states, while identifying limitations and potential improvements

III. Develop a generic method capable of measuring the relevant emotional states. The method should provide a continuous measure of emotion in real-time, while also requiring as minimal calibration as possible. Perform a detailed comparison / validation in a relevant (to this thesis) case study with the previously identified methodologies

IV. Study the correlation between emotional states and the various facets of user experience (immersion, tension, flow, etc.), as well as the impact of static emotionally driven gameplay adaptation schemes

V. Propose a grounded methodology to: a) automatically associate the emotional reactions to the eliciting interaction events, and b) compile the observed emotional reactions into players' affective reaction models

VI. Define an adaptation scheme to leverage the user collated data in order to reinforce a set of desired emotional states and patterns

VII. Test the created models / adaptation scheme to assess their general emotional elicitation capabilities on a wide range of emotional states.

Since most of these objectives have very strong dependencies on the previous ones, throughout this thesis we tackled each of them in a sequential fashion. In the following paragraphs we will provide a condensed view of the achieved results in each chapter, tying them to the objectives they aimed at completing. We also comment on the observed limitations, which lead us to discussing the potential for future work in each of this thesis' addressed topics. Since in Chapter I we merely introduced the thesis motivation, hypothesis and proposed objectives, a summary of its findings is trivially omitted.

In Chapter I, we also tackled our thesis' first objective – to propose a conceptual architecture for the creation of emotionally adaptive systems. This architecture was targeted at being a system that could interface with any interactive application, as long as a proper interface was exposed. To increase its number of potential use cases, it was also designed so as to not rely on a specific input type, only specifying that emotional states should be numeric data with a variable number of

dimensions (hence the choice of Russell's AV space) and game events as nominal data with an associated timestamp. This way we assured emotional reactions could be both quantified and tied to the triggering events. Since this was a conceptual description of the architecture, no drawbacks existed *per se*. From a design standpoint, the biggest drawback is that while the core components of the architecture can be used, the interfacing components cannot as the application interface and emotional recognition method may have to change due to study or logistic constraints. However, for the system to adapt to the highest possible number of scenarios, this was a necessary design pattern.

In Chapter II we presented an introduction to the psychophysiological mechanisms relevant to this thesis' work and provided proper definitions of the employed user experience and affective computing jargon, such as immersion, flow and emotional regulation. We also studied the effects of these concepts in terms of user experience, the potential (positive or otherwise) impact on people's lives and closely examined the link between affective experiences (emotions) and user experience dimensions (immersion, flow, etc.). While this did not achieve any of our thesis objectives or posed a significant scientific contribution, it was a necessary briefing of human psychophysiological activity and a contextualisation for the usage of emotions as catalysts of improved experiences.

Having both defined how our ideal emotionally adaptive system should work and clarified the necessary psychophysiological concepts, our focus then turned towards describing each of its components. Since emotional detection was not only a critical component of our system, but also a highly complex task with no major consensus by the scientific community we expanded on our original objective and strived to provide a comprehensive analysis on the matter. Thus, Chapter III presented a series of development guidelines on how to develop emotional detection systems for both on and offline applications.

We started by presenting a theoretical/data-driven hybrid, multi-layered method to interpret selected psychophysiological measures in terms of the arousal and valence affect dimensions. The exhibited results show that we were able to successfully address the recurring emotional scaling and participant physiological activation function normalisation issues present in the literature through a combination of careful experimental design and a mixture of linear and low-degree polynomial regression models. Furthermore, this regression process allowed us to keep the system complexity relatively low, as well as generate humanly interpretable models. Since the regression models properly scale the ratings across individuals, this also allowed us to

generalise the second classification layer. This means that upon an initial calibration of the regression models, the method is participant-independent.

Given that we were able to build this system in a participant-independent fashion, it proved able to adequately generalise within the considered affective experiences and population. The system's biggest limitation is that despite this, subsequent tests on a larger population are needed for a strong generalisation proof outside of our controlled experimental conditions and demographic.

To conclude our analysis on emotional detection systems, we also created a simplified version of our previous method for usage in real-time or low fidelity systems. The validation tests using a manual approach based on the literature's accepted good practices showed that both the online and the offline approaches were on track and managed to improve on the manual method's results, thus presenting viable, improved alternatives.

Within the overarching scope of this thesis, Chapter III addressed objectives II and III. Firstly, it defined which existing state of the art methodologies were best suited for quantitatively and non-intrusively measure emotional states, while at the same time, identifying limitations and potential improvements. Secondly, we presented a generic method capable of measuring the relevant emotional states in real-time and with minimal calibration.

Having established how to capture players' emotional states in a real-time fashion, in Chapter IV we turned our attention towards creating an emotionally-driven biofeedback game. This allowed us to: *1)* test the effectiveness of emotionally-adaptive games and their impact on players' emotional states (i.e. assess whether they are able to significantly alter players' affect), and *2)* study the correlations between players' emotional states and various dimensions of user experience, such as immersion, tension, and flow, amongst others (thesis objective IV). Ultimately, we showed that player experience can be influenced in a statistically significant way by the integration of basic (static) biofeedback mechanisms.

We also aimed at improving the rather dispersed studies in biofeedback techniques by providing a comparative analysis on them and defining a consistent classification scheme for biofeedback games. Furthermore, by leveraging the collection of a wide body of data from our study, we were also able to do this analysis in a more thorough, detailed manner, spanning multiple points of view (objective vs subjective); a study that from our analysis was missing in the literature.

Overall, the study presented in Chapter IV showed evidence in favour of augmenting game mechanics with affective physiological data. Given that the presented "dumb" (static) emotional regulation mechanisms were able to elicit significant changes in players' affective and user experience, this would suggest "smarter" (dynamic) mechanisms would only enhance these results, which strengthens our thesis hypothesis. However, we focused on a specific game genre to perform a deeper analysis following previous investigations and thus our work is limited in its applicability to other game genres.

Additionally, although we explored various game mechanics on two different indirect biofeedback types, different choices with regard to game mechanics and biofeedback would most likely have resulted in different player experience results. Furthermore, despite providing enough time for players to clearly notice differences in the gameplay experience, the budgeted playtime was still not sufficient to account for novelty and habituation effects to this new technology. To provide a more complete investigation of this technology, these effects should be evaluated in future long-term studies.

From a perspective related to commercial biofeedback-augmented games, this technology's most significant limitations are: *1)* the game designer's willingness to incorporate physiological interaction into their core gameplay mechanics, and *2)* the physiological sensors' proper calibration. Usually, the availability of the necessary hardware is mentioned as the primary challenge of using this technology, but we feel that this argument is losing momentum with the introduction of low-cost solutions like BITalino and the expression of interest in such technologies by industry giants such as Valve. Regarding sensor calibration, a potential solution would be to borrow the attributes of natural interaction solutions (e.g., Kinect and WiiMote) and gamify the calibration process, effectively masking it within the game's initial tutorial sections (Flatla et al., 2011).

Having collected players' emotional states and game event logs for all three gaming conditions presented in Chapter IV, we then turned our attention towards modelling players' emotional reactions as closely as possible. In this effort, we first developed a tool for the automatic annotation of emotional reactions to digital stimuli. This work, presented in Chapter V, was also meant to contribute to a wider accessibility of emotional response studies by automating the annotation process and also removing the necessity of developing standalone emotional state detection systems. We expected these to reduce the traditional human error associated with the subjective

manual annotation process, thus contributing to a standardisation and comparability of obtained results.

Once again recalling our thesis objectives, the main contribution of Chapter V towards dynamic affective experiences is the emotional reaction extraction method, which corresponds to part A of objective V: *"Propose a grounded methodology to… automatically associate the emotional reactions to the eliciting interaction events"*.

Armed with players' emotional reactions to Vanish's game events, in Chapter VI we attempted to create models of player affect. Our first attempt was based on a purely feature-driven (black box) approach. Driven by the lack of a continuous output, the difficulty in model updating and sparse nature of our dataset, we then shifted towards a more domain knowledge – but still data driven – clustering method. This method attempted to first approximate players' individual reactions to each game event via a matrix of polynomial regression models, which were then used to create a distance matrix between player pairs. Using a hierarchical clustering algorithm we then obtained a set of player clusters on which we could map both existing and new players via fuzzy memberships. This process overcame several of our previous approach's limitations, enabling continuous output, seamless incorporation of new training data and the ability to handle sparse data by making player approximation models and using the whole player population's dataset to create the clusters. In terms of our thesis' objectives, in Chapter VI we addressed part B of objective V: "*(To) propose a grounded methodology to… compile the observed emotional reactions into players' affective reaction models"*.

Despite the limitations of the two aforementioned approaches, both were shown to be reliable and generic methods for modelling players' affective reactions. Do note that games that do not meaningfully elicit any physiological alterations on the player would result in, albeit correct, null response models. In theory, these models can be applied to any form of digital stimuli, although the same stimuli potency issue may be present. Having said this, we would also like to point out that despite focusing on physiologically interpreted emotional state data, this is not a requirement for our method as any numerical representation of the player's emotional state will suffice. The same is true for the arousal and valence dimensions of emotion. As such, should the chosen stimuli not provide any meaningful physiological alterations, it is permissible that the emotional detection system be replaced in favour of other metrics, such as for example, facial recognition, body posture, text sentiment analysis, or context-based interaction analysis.

The models proposed in Chapter VI are thus not only useful for affective games. One of the most immediate contributions posed by them is the potential to both accelerate and increase the objectivity of the (affective) game design process. For example, besides using the models to drive the adaptation of the gameplay experience automatically in real-time, the affective player models can be used during *playtesting* phases to inform game designers on which stimuli are most effective at eliciting a set of target emotional states.

It was with this objective that, in Chapter VII, we used the created models as the key driver component for a symbolic simulator based on our case study game; Vanish. This symbolic simulator ran a highly abstracted version of Vanish and was steered by the emotional reactions predicted by players' affective reaction models. In the final chapter of this thesis we used this simulator to explore two main (potential) applications of our player reaction models: *1)* using the models to run offline simulations that allowed us to identify the game's optimal configuration for a specific emotional state, and *2)* using these same models to drive the game's playout so as to provide real-time, player-specific adaptation. These premises were tested in two experiments that used no gameplay adaptation (OBO) and periodic gameplay adaptation based on the optimal event sets predicted by player models (ORB).

Regarding our first question, through the OBO experiment we saw that it is indeed possible to optimise the game's configuration towards specific emotional states. However, this comes with a rather high computational price in sampling the whole game configuration space. Since it is a static optimization, is also unable to elicit emotional patterns. This limitation was however overcome by the use of a dynamic adaptation mechanism in the ORB experiment, which boasted improved results across the board as well as increased stability.

Another key contribution present in this final chapter was the developed simulation platform itself, which was intended as a generic tool for games and affective computing research in general. With its development, we expect that in the future it can provide game designers and affective computing researchers with a common, stable platform for automating, reproducing and validating research results.

In terms of the proposed thesis objectives, in Chapter VII we addressed the thesis' two final objectives: VI: *"Define an adaptation scheme to leverage the user collated data in order to reinforce a set of desired emotional states and patterns"* and VII: *"Test the created models / adaptation scheme to assess their general emotional elicitation capabilities on a wide range of emotional states"*. The formed refers, in

general, to ORB's simulation flow and was addressed in the simulator's description at the beginning of this chapter. The latter essentially stated that we wished to test the effectiveness of the created models and is the reason behind the elaboration of both the OBO and ORB experiments.

In conclusion, throughout this thesis we have successfully described an adaptive emotion regulation framework (our Emotion Engine) and implemented all of its fundamental components. Throughout this process we have improved on the current state of the art for emotional detection, proposing systems for both online and offline applications. We have also gone a step further than what was originally planned and developed a tool for the standardisation of physiological and emotional data in affective computing studies. Most importantly, we have shown that static biofeedback adaptation mechanisms produce (statistically) significant improvements to user experience. Following these results we studied the feasibility of leveraging players' emotional reactions to build models of user affect that could be used to driven more intelligent (dynamic) biofeedback adaptation mechanisms. The obtained results show that all of these can be achieved with significant accuracy ratings and produce substantial results when used to elicit specific emotional states – be they static ones or emotional patterns.

Despite the aforementioned limitations of this technology – habituation effects, hardware and logistic constraints and inability to fully counterbalance a would-be lack of potency from the selected stimuli – we believe emotional regulation is a powerful tool with tremendous potential that had yet to be fully explored. As discussed in Chapter II, emotions are complex psychophysiological experiences of an individual, influenced by a state of mind, which arises as the result of interacting biochemical reactions and environmental interactions. As they occur at a deep and sometimes instinctual or subconscious level, emotions influence humans in a very meaningful and critical way, often overriding even rational thought. This is one of the reasons why they are one of the main players in the formation of thoughts and, consequently, ideas. Therefore, we believe that the usage of the players' emotions as catalysts for a reactive system embedded in a game engine is a viable and effective means to improve the overall user experience. As it stands, the proposed work suggests an adaptable system that, being able to flourish in a hostile, low SNR (i.e. multiple parallel events and stimuli) environment such as an interactive virtual world has its general transferability to virtually any other area assured. In theory, this work is applicable in any system that has a lasting and/or complex interaction scheme with its users, thus posing a significant advance in

the fields of affective computing, artificial intelligence and human-machine interaction.

REFERENCES

REFERENCES

Abrilian, S., Devillers, L., Buisine, S., & Martin., J. C. (2005). EmoTV1: Annotation of Real- life Emotions for the Specification of Multimodal Affective Interfaces. In HCI International. Las Vegas, USA.

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. MIS Quarterly, 24(4), 665–694. Retrieved from http://www.jstor.org/stable/10.2307/3250951

Ambinder, M. (2011). Biofeedback in Gameplay: How Valve Measures Physiology to Enhance Gaming Experience. In Game Developers Conference.

Baños, R. M., Botella, C., Alcañiz, M., Liaño, V., Guerrero, B., & Rey, B. (2004). Immersion and emotion: their impact on the sense of presence. Cyberpsychology & Behavior : The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society, 7(6), 734–41. doi:10.1089/cpb.2004.7.734

Barakova, E. I., Spink, A. S., Boris de Ruyter, L., & Noldus, P. J. J. (2013). Trends in measuring human behavior and interaction. Personal and Ubiquitous Computing, 17(1), 1–2. doi:10.1007/s00779-011-0478-x

Bernhaupt, R., Boldt, A., & Mirlacher, T. (2007). Using emotion in games: emotional flowers. In Proceedings of the international conference on Advances in computer entertainment technology (ACE) (pp. 41–48). doi:10.1145/1255047.1255056

Bersak, D., McDarby, G., Augenblick, N., McDarby, P., McDonnell, D., McDonald, B., & Karkun, R. (2001). Intelligent biofeedback using an immersive competitive environment.

Blanchard, E. B., Eisele, G., Vollmer, A., Payne, A., Gordon, M., Cornish, P., & Gilmore, L. (1996). Controlled evaluation of thermal biofeedback in treatment of elevated blood pressure in unmedicated mild hypertension. Biofeedback and Self-Regulation, 21(2), 167–190.

Bourg, S. (2004). AI for Game Developers. O'Reilly & Associates.

Brown, E., & Cairns, P. (2004). A grounded investigation of game immersion. In Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04 (p. 1297). New York, New York, USA: ACM Press. doi:10.1145/985921.986048

Bryant, M. A. M. (1991). Biofeedback in the treatment of a selected dysphagic patient. Dysphagia, 6(2), 140–144.

C., C., & H., M. (2009). Empirically Building and Evaluating a Probabilistic Model of User Affect. User Modeling and User-Adapted Interaction, 19, 267–303.

Caldognetto, E. M., Poggi, I., Cosi, P., Cavicchio, F., & Merola, G. (2004). Multimodal Score: an ANVILTM Based Annotation Scheme for Multimodal Audio-Video Analysis. In International Conference on Language Resources and Evaluation Workshop on Multimodal Corpora (pp. 29–33).

Calleja, G. (2011). In-Game: From Immersion to Incorporation (1st ed.). The MIT Press.

Cannon, W. B. (1929). Bodily changes in pain, hunger, fear, and rage. New York: Appleton-Century-Crofts.

Cavazza, M., Pizzi, D., Charles, F., Vogt, T., & André, E. (2009). Emotional input for character-based interactive storytelling. In Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1 (pp. 313–320). International Foundation for Autonomous Agents and Multiagent Systems.

Chanel, G., Kronegg, J., Grandjean, D., & Pun, T. (2006). Emotion Assessment : Arousal Evaluation Using EEG's and Peripheral Physiological Signals. In Proc. Int. Workshop on Multimedia Content Representation, Classification and Security (pp. 530–537). Springer.

CNBC. (2010). Video Game Sales Drop 6% in 2010. Retrieved from www.cnbc.com/id/41062675/Video_Game_Sales_Drop_6_in_2010_Second_Year_of_Declines

Cox, A. L., Cairns, P., Berthouze, N., & Jennett, C. (2006). The Use of Eyetracking for Measuring Immersion. workshop on What have eye movements told us so far, and what is next? In Twenty-Eighth Annual Meeting of the Cognitive Science Society (CogSci2006) (p. N/A). Vancouver, Canada.

Csikszentmihalyi, M. (1988). The flow experience and its significance for human psychology. In M. Csikszentmihalyi & I. S. Csikszentmihalyi (Eds.), Optimal experience Psychological studies of flow in consciousness (pp. 15–35). Cambridge University Press.

Csíkszentmihályi, M. (2008). Flow: The Psychology of Optimal Experience (p. 336). HarperCollins.

Csikszentmihalyi, M., & Rathunde, K. (1993). "The measurement of flow in everyday life: Towards a theory of emergent motivation" in Developmental perspectives on motivation. Nebraska symposium on motivation. (J. E. Jacobs, Ed.). Lincoln: University of Nebraska Press.

Damasio, A. (1994a). Descartes' error: Emotion, reason and the human brain. NewYork:Gosset/Putnam Press.

Damasio, A. (1994b). Descartes' Error: Emotion, Reason, and the Human Brain (p. 336). Penguin Books.

Dekker, A., & Champion, E. (2007). Please biofeed the zombies: enhancing the gameplay and display of a horror game using biofeedback. In Situated Play, Proceedings of the Digital Games Research Association (DiGRA) Conference (pp. 550–558).

Dodd, J., & Role, L. W. (1997). The autonomic nervous system. Principles of neural science. (E. R. Kandel, J. H. Schwartz, & T. M. Jessel, Eds.) (3rd ed.). Appleton & Lange.

Dong, Q., Li, Y., Hu, B., Liu, Q., Li, X., & Liu, L. (2010). A Solution on Ubiquitous EEG-based Biofeedback Music Therapy. In IEEE 5th International Conference on Pervasive Computing and Applications (ICPCA) (pp. 32–37). doi:http://dx.doi.org/10.1109/ICPCA.2010.5704071

Douglas, Y. (2000). The pleasure principle: immersion, engagement, flow. In Proceedings of the eleventh ACM on Hypertext and hypermedia (pp. 153–160).

Drachen, A., Nacke, L. E., Yannakakis, G., & Pedersen, A. L. (2010). Correlation between Heart Rate, Electrodermal Activity and Player Experience in First-Person Shooter Games. In Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games (pp. 49–54). ACM.

Ekman, P. (2003). Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life (1st ed.). Times Books.

"Emotion." (2010). In Oxford Dictionaries. Oxford University Press.

Entertainment Software Rating Board: Video Game Industry Statistics. (2012). Retrieved April 19, 2012, from www.esrb.org/about/video-game-industry-statistics.jsp

Ermi, L., & Mäyrä, F. (2005). Fundamental components of the gameplay experience: Analysing immersion. In Digital Games Research Association Conference: Changing Views - Worlds in Play.

Etheredge, M., Lopes, R., & Bidarra, R. (2013). A Generic Method for Classification of Player Behavior. In IDPv2 2013 - Workshop on Artificial Intelligence in the Game Design Process.

Figueiredo, R., & Paiva, A. (2010). "I want to slay that dragon" - Influencing Choice in Interactive Storytelling. In Digital Interactive Storytelling.

Flatla, D. R., Gutwin, C., Nacke, L. E., Bateman, S., & Mandryk, R. L. (2011). Calibration games: making calibration tasks enjoyable by adding motivating game elements. In Proceedings of the 24th annual ACM symposium on User interface software and technology (pp. 403–412). doi:10.1145/2047196.2047248

Gilleade, K. M., Dix, A., & Allanson, J. (2005). Affective Videogames and Modes of Affective Gaming : Assist Me , Challenge Me , Emote Me. In DIGRA - Digital Games Research Association (pp. 1–7).

Gilroy, S., & Porteous, J. (2012). Exploring passive user interaction for adaptive narratives. Intelligent User Interaction, 119–128.

Gilroy, S. W., Cavazza, M., & Benayoun, M. (2009). Using affective trajectories to describe states of flow in interactive art. In Proceedings of the International Conference on Advances in Computer Enterntainment Technology - ACE '09 (p. 165). New York, New York, USA: ACM Press. doi:10.1145/1690388.1690416

Gow, J., Cairns, P., Colton, S., Miller, P., & Baumgarten, R. (2010). Capturing Player Experience with Post-Game Commentaries. In Proceedings of the 3rd Annual International Conference Computer Games Multimedia Allied Technology. Citeseer.

Gratch, J., & Marsella, S. (2005). Evaluating a Computational Model of Emotion. In Autonomous Agents and Multi-Agent Systems (pp. 23–43).

Gross, J. J., & Thompson, R. A. (2009). Emotion regulation: Conceptual foundations in Handbook of emotion regulation. (J. J. Gross, Ed.) (1st ed.). New York: Guilford Press.

Gunes, H., & Pantic, M. (2010). Automatic, Dimensional and Continuous Emotion Recognition. International Journal of Synthetic Emotions, 1(1), 68–99.

Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004). Emotion recognition using bio-sensors: First steps towards an automatic system. Affective Dialogue Systems.

Hazlett, R. (2006). Measuring Emotional Valence during Interactive Experiences : Boys at Video Game Play. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 1023–1026).

Hazlett, R., & Benedek, J. (2007). Measuring emotional valence to understand the user's experience of software. International Journal of Human-Computer Studies, 65(4), 306–314. doi:10.1016/j.ijhcs.2006.11.005

Hellquist, P., Levine, K., & Schyman, G. (2007). BioShock. 2K Games & Feral Interactive (MacOS X).

Hepach, R., Kliemann, D., Grüneisen, S., Heekeren, H. R., & Dziobek, I. (2011). Conceptualizing emotions along the dimensions of valence, arousal, and communicative frequency - implications for social-cognitive tests and training tools. Frontiers in Psychology, 2(October), 266. doi:10.3389/fpsyg.2011.00266

Hjelm, S. I., & Browall, C. (2000). Brainball–Using brain activity for cool competition. In Proceedings of NordiCHI (pp. 177–188).

Holmgård, C., Togelius, J., & Yannakakis, G. (2013). Decision Making Styles as Deviation from Rational Action A Super Mario Case Study. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment.

Huang, H., Ingalls, T., Olson, L., Ganley, K., Rikakis, T., & He, J. (2005). Interactive multimodal biofeedback for task-oriented neural rehabilitation. In 27th Annual International Conference of the Engineering in Medicine and Biology Society (IEEE-EMBS) (pp. 2547–2550).

Hudlicka, E. (2009). Affective Game Engines : Motivation and Requirements. In International Conference on Foundations of Digital Games. Orlando, Florida, USA.

IJsselsteijn, W. A., Poels, K., & De Kort, Y. (2008). The game experience questionnaire: Development of a self-report measure to assess player experiences of digital games: FUGA technical report, Deliverable 3.3. Eindhoven.

Ijsseltein, W. A., Poels, K., & De Kort, Y. (2008). The game experience questionnaire: Development of a self-report measure to assess player experiences of digital games: FUGA technical report, Deliverable 3.3.

Jennett, C., Cox, A. L., & Cairns, P. (2008). Being " In the Game ." In Conference Proceedings of the Philosophy of Computer Games (pp. 210–227).

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. International Journal of Human-Computer Studies, 66(9), 641–661. doi:10.1016/j.ijhcs.2008.04.004

Kikuchi, K. (2012). Fatal Frame II: Crimson Butterfly. Tecmo, Ubisoft, Microsoft Game Studios & Nintendo.

Kim, J., Bee, N., Wagner, J., & André, E. (2004). Emote to win: Affective interactions with a computer game agent. In GI Jahrestagung: (1) (pp. 159–164).

Kleinginna, P., & Kleinginna, A. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. Motivation and Emotion. doi:5:345-379

Kuikkaniemi, K., Laitinen, T., & Turpeinen, M. (2010). The influence of implicit and explicit biofeedback in first-person shooter games. In Proceedings of the 28th international conference on Human factors in computing systems (pp. 859–868).

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system (IAPS).

Lankes, M., Hochleitner, W., Hochleitner, C., & Lehner, N. (2012). Control vs. complexity in games: comparing arousal in 2D game prototypes. In Proceedings of the 4th International Conference on Fun and Games (pp. 101–104).

Lazarus, R., & Lazarus, B. (1994). Passion and Reason: Making Sense of Our Emotions. Oxford University Press.

Lazzaro, N. (2005). Why We Play Games : Four Keys to More Emotion Without Story. Design, 18, 1–8. doi:10.1111/j.1464-410X.2004.04896.x

Leite, I., Pereira, A., Mascarenhas, S., Castellano, G., Martinho, C., Prada, R., & Paiva, A. (2010). Closing the Loop: From Affect Recognition to Empathic Interaction. In 3rd Int. Workshop on Affect Interaction in Natural Environments.

Leon, E., Clarke, G., Callaghan, V., & Sepulveda, F. (2007). A user-independent real-time emotion recognition system for software agents in domestic environments. Engineering Applications of Artificial Intelligence, 20(3), 337–345. doi:10.1016/j.engappai.2006.06.001

Levillain, F., Orero, J. O., Rifqi, M., & Bouchon-Meunier, B. (2010). Characterizing Player's Experience From Physiological Signals Using Fuzzy Decision Trees. In IEEE Symposium on Computational Intelligence and Games (CIG) (pp. 75–82).

Loveridge, S. (2014). Sony confirms biometric sensors were tested for the PS4 controller. TrustedReviews. Retrieved March 31, 2014, from http://www.trustedreviews.com/news/sony-confirms-biometric-sensors-were-tested-for-the-ps4-controller

Mandryk, R., & Atkins, M. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human-Computer Studies, 65(4), 329–347. doi:10.1016/j.ijhcs.2006.11.011

Mandryk, R. L. (2005). Modeling User Emotion in Interactive Play Environments: A Fuzzy Physiological Approach.

Mandryk, R. L., Dielschneider, S., Kalyn, M. R., Bertram, C. P., Gaetz, M., Doucette, A., … Keiver, K. (2013). Games as neurofeedback training for children with FASD. In Proceedings of the 12th International Conference on Interaction Design and Children (IDC '13) (pp. 165–172). doi:10.1145/2485760.2485762

Marshall, J., Rowland, D., Egglestone, S. R., Benford, S., Walker, B., & McAuley, D. (2011). Breath control of amusement rides. In Proceedings of the SIGCHI ACM Conference on Human Factors in Computing Systems (pp. 73–82).

Martinez, H. P., Garbarino, M., & Yannakakis, G. N. (2011). Generic Physiological Features as Predictors of Player Experience. In Proceedings of the 2011 Affective Computing and Intelligent Interaction Conference (pp. 267–276).

Matias Kivikangas, J., Nacke, L., & Ravaja, N. (2011). Developing a triangulation system for digital game events, observational video, and psychophysiological data to study emotional responses to a virtual character. Entertainment Computing, 2(1), 11–16. doi:10.1016/j.entcom.2011.03.006

Maybury, M. T., & Kipp., M. (2012). Multimedia Annotation, Querying and Analysis in ANVIL. In Multimedia Information Extraction: Advances in Video, Audio, and Imagery Analysis for Search, Data Mining, Surveillance, and Authoring. doi:10.1002/9781118219546

Minsky, M. (1980). "Telepresence." MIT Press Journals, 45–51.

Moreira, V. H. V. G. (2010). BioStories Geração de Conteúdos Multimédia Dinâmicos Mediante Informação Biométrica da Audiência.

Nacke, L. E. (2013). An introduction to physiological player metrics for evaluating games. In Game Analytics (pp. 585–619). Springer London.

Nacke, L. E., Kalyn, M., Lough, C., & Mandryk, R. L. (2011). Biofeedback Game Design : Using Direct and Indirect Physiological Control to Enhance Game Interaction. In Proceedings of the 2011 annual conference on Human factors in computing systems (pp. 103–112). ACM.

Nacke, L. E., Stellmach, S., & Lindley, C. a. (2010). Electroencephalographic Assessment of Player Experience: A Pilot Study in Affective Ludology. Simulation & Gaming. doi:10.1177/1046878110378140

Nacke, L., & Lindley, C. A. (2008). Boredom, Immersion, Flow - A Pilot Study Investigating Player Experience. In Conference on Game and Entertainment Technologies.

Nacke, L., & Lindley, C. A. (2008). Flow and immersion in first-person shooters: measuring the player's gameplay experience. In Proceedings of the 2008 Conference on Future Play: Research, Play, Share (pp. 81–88). ACM.

Nakazawa, K., Yamaoka, A., Ito, M., & Owaku, H. (2003). Silent Hill 3. Konami Computer Entertainment Tokyo.

Narcisse, E. (2013). Kinect 2.0 Sees Your Face, Muscles and Soul. Maybe Not That Last One. Kotaku.

Nasoz, F., Lisetti, C. L., Alvarez, K., & Finkelstein, N. (2003). Emotion Recognition from Physiological Signals for User Modeling of Affect. In Proceedings of the 3rd Workshop on Affective and Attitude User Modelling. Pittsburgh, PA, USA.

Negini, F., Mandryk, R. L., & Stanley, K. (2014). Using Affective State to Adapt Characters, NPCs, and the Environment in a First-Person Shooter Game. In IEEE Games, Entertainment, and Media (p. Too appear). Toronto, Canada.

Netter, F. H. (1997). Atlas of Human Anatomy (2nd ed., p. 525). Rittenhouse Book Distributors, Inc.

Noback, C. R., Ruggiero, D. A., Demarest, R. J., & Strominger, N. L. (2005). The Human Nervous System: Structure and Function (6th ed., p. 416). Humana Press.

Ortony, A., Clore, G., & Collins, A. (1990). The Cognitive Structure of Emotions. Cambridge University Press.

Parnandi, A., & Gutierrez-Osuna, R. (2014). A comparative study of game mechanics and control laws for an adaptive physiological game. Journal on Multimodal User Interfaces. doi:10.1007/s12193-014-0159-y

Parnandi, A., Son, Y., & Gutierrez-Osuna, R. (2013a). A Control-Theoretic Approach to Adaptive Physiological Games. In Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on (pp. 7–12). IEEE. doi:10.1109/ACII.2013.8

Parnandi, A., Son, Y., & Gutierrez-Osuna, R. (2013b). A Control-Theoretic Approach to Adaptive Physiological Games. In IEEE Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 7– 12).

Pedersen, C., Togelius, J., & Yannakakis, G. N. (2009). Modeling Player Experience for Content Creation. Computational Intelligence and AI in Games, 2(1), 121–133.

Picard, R. W. (1995). Affective Computing MIT Technical Report #321.

Pigna, K. (2009). Sony Patents “Emotion Detecting” PS3 Technology. 1Up.com. Retrieved from http://www.1up.com/news/sony-patents-emotion-detecting-ps3

Plutchik, R. (1980). A General Psychoevolutionary Theory of Emotion. Emotion: Theory, Research, and Experience, 1(1), 3–33.

Pope, A. T., Stephens, C. L., & Gilleade, K. (2014). Biocybernetic Adaptation as Biofeedback Training Method. In Advances in Physiological Computing (pp. 91–115).

Rani, P., Sarkar, N., & Liu, C. (2005). Maintaining optimal challenge in computer games through real-time physiological feedback. In Proceedings of the 11th International Conference on Human Computer Interaction (pp. 184–192).

Rau, P.-L. P., Peng, S.-Y., & Yang, C.-C. (2006). Time Distortion for Expert and Novice Online Game Players. CyberPsychology & Behavior, 9(4), 396–403.

Ravaja, N., Salminen, M., Holopainen, J., Saari, T., Laarni, J., & Jarvinen, A. (2004). Emotional response patterns and sense of presence during video games: Potential criterion variables for game design. In Proceedings of the third Nordic conference on Human-computer interaction (pp. 339–347). doi:10.1145/1028014.1028068

Reynolds, E. (2013). Nevermind. Retrieved from http://www.nevermindgame.com/Nevermind_Game/Nevermind__The_Game.html

Riva, G., Gaggioli, A., Pallavicini, F., Algeri, D., Gorini, A., & Repetto, C. (2010). Ubiquitous Health for the Treatment of Generalized Anxiety Disorders. In UbiComp ‘10. Copenhagen, Denmark.

Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., … Alcañiz, M. (2007). Affective Interactions Using Virtual Reality: The Link between Presence and Emotions. CyberPsychology & Behavior, 10(1), 45–56. doi:10.1089/cpb.2006.9993

Rocchi, L., Benocci, M., Farella, E., Benini, L., & Chiari, L. (2008). Validation of a wireless portable biofeedback system for balance control: preliminary results. In IEEE Second International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth (pp. 254–257).

Rokach, L. (2005). Ensemble-based classifiers. Artificial Intelligence Review, 33(1-2), 1–39.

Russel, J. A. (1980a). A Circumplex Model of Affect. Journal of Personality and Social Psychology, 39(6), 1161–1178.

Russel, J. A. (1980b). A Circumplex Model of Affect. Personality and Social Psychology, 39(6), 1161–1178.

S, A., E, W., P, N., J, H., L, M., H, S., & H., K. (2005). Regression analysis utilizing subjective evaluation of emotional experience in PET studies on emotions. Brain Res Brain Res Protoc., 15(3), 142–154.

Schaefer, A., Nils, F., Sanchez, X., & Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films : A new tool for emotion researchers. Cognition and Emotion, 24(7), 1153–1172.

Schofield, G., Robbins, B., Ellis, W., Remender, R., Johnston, A., & Graves, J. (2008). Dead Space. Electronic Arts.

Sennett, R. (1993). Authority. W. W. Norton & Company.

Shaker, N., Yannakakis, G., & Togelius, J. (2010). Towards automatic personalized content generation for platform games. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (pp. 63–68).

Slater, M. (2003). A Note on Presence Terminology. Presence Connect, 3(3).

Slater, M., & Usoh, M. (1994). Body centred interaction in immersive virtual environments. Artificial Life and Virtual Reality, 125–148.

Stearns, C., & Stearns, P. (1989). Anger: The Struggle for Emotional Control in America's History. America's History. The University of Chicago Press.

Stepp, C. E., Britton, D., Chang, C., Merati, A. L., & Matsuoka, Y. (2011). Feasibility of game-based electromyographic biofeedback for dysphagia rehabilitation. In 5th International IEEE/EMBS Conference on Neural Engineering (NER) (pp. 233–236).

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). Psychophysiological recording (2nd ed.). New York: Oxford University Press.

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning. Psychological Methods, 4(14), 323–348.

Suzuki, R., & Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics, 22(12), 1540–1542.

Tijs, T. J. W. (2006). Quantifying Immersion in Games by Analyzing Eye Movements. Department of Computer and Systems Science, Royal Institute of Technology, Stockholm (pp. 1–4). Retrieved from

http://www.benschweitzer.org/WORK/game_heuristics/Quantifying_immersion_in_games_by_analyzing_eye_movements.pdf

Tognetti, S., Garbarino, M., Bonarini, A., & Matteucci, M. (2010). Modeling Enjoyment Preference from Physiological Responses in a Car Racing Game. In IEEE Conference on Computational Intelligence and Games (pp. 321–328).

Turner, P. (2010). The anatomy of engagement. In Proceedings of the 28th Annual European Conference on Cognitive Ergonomics (Vol. 44, pp. 25–27). Retrieved from http://dl.acm.org/citation.cfm?id=1962315

Vachiratamporn, V., Legaspi, R., Moriyama, K., & Numao, M. (2013). Towards the Design of Affective Survival Horror Games: An Investigation on Player Affect. In Affective Computing and Intelligent Interaction (ACII) (pp. 576–581).

Vinhas, V., Silva, D., Oliveira, E., & Reis, L. (2009). Biometric Emotion Assessment and Feedback in an Immersive Digital Environment. Social Robotics, 307–317.

Vorderer, P., & Bryant, J. (2006). Playing Video Games: Motives, Responses, and Consequences (pp. 183–184). Lawrence Erlbaum Associates.

Wang, N., & Marsella, S. (2006). Introducing EVG: An Emotion Evoking Game. In Intelligent Virtual Agents (pp. 282–291).

Wood, R. T. A., Griffiths, M. D., & Parke, A. (2007). Experiences of Time Loss among Videogame Players: An Empirical Study. CyberPsychology & Behavior, 10(1), 38–44.

Yang, Y.-H. (2008). A Regression Approach to Music Emotion Recognition. IEEE Transactions on Audio, Speech, and Language Processing, 16(2), 448–457.

Yannakakis, G. N., & Togelius, J. (2011). Experience-driven Procedural ContentGeneration. Transactions on Affective Computing, 2(3), 147-161.

A P P E N D I X E S

PHYSIOLOGIC DATA MAPPING ALGORITHMS

Algorithm A. Rules for mapping the regressed physiological metrics into the AV dimensions in the grounded approach.
The following algorithm was applied on top of the regressed physiological metrics to convert regressed skin conductance (RSC), regressed heart rate (RHR) and regressed electromyography ($REMG_{zyg}$ and $REMG_{corr}$) into arousal and valence.

Algorithm A: Grounded AV Approach

Inputs:
Regressed arousal indicators
$\{ RSC, RHR_A \}$
Regressed valence indicators
$\{ REMG_{ZYG}, REMG_{CORR}, RHR_V \}$

Parameters:
Voting weights for each physiological correlation
$\{ W_{HRa}, W_{SC}, W_{EMG}, W_{HRv} \}$
Electromyographic valence prediction EVP
Activity magnitude function M
Total EMG-related activity EMG_{ACT}
Minimal EMG activation EMG_{MIN} (default: 0.5)
Maximal EMG weight $WEMG_{MAX}$ (default: 0.8)

Output:
Arousal prediction A
Valence prediction V

if $RSC = \emptyset$ **then**
  $A \leftarrow RHR_A$
elif $RHR = \emptyset$ **then**
  $A \leftarrow RSC$
elif $RHR_A \geq 7.0$ **and** $RSC \leq 7.0$ **then**
  $W_{HRa} \leftarrow 1/3$
  $W_{SC} \leftarrow 1 - W_{HR}$
  $A \leftarrow W_{HRa} * RHR_A + W_{SC} * RSC$
else
  if $| RHR_A - RSC | > 1.5$ **then**
    $W_{HRa} \leftarrow 0$
    $W_{SC} \leftarrow 1$
  else
    $W_{HRa} \leftarrow 1 - (| RHR_A - RSC | / 1.5)$
    $W_{SC} \leftarrow 2 - (| RHR_A - RSC | / 1.5)$
  $A \leftarrow (W_{HRa} \times RHR_A + W_{SC} \times RSC) / (W_{HRa} + W_{SC})$
$EMG_{ACT} \leftarrow M(REMG_{ZYG}) + M(REMG_{CORR})$
$EVP \leftarrow ((M(REMG_{ZYG}) / EMG_{ACT})) \times REMG_{ZYG} + ((M(REMG_{CORR}) / EMG_{ACT})) \times REMG_{CORR}$

19 $W_{EMG} \leftarrow$ (MINIMAL(| $EVP - 5.0$ |, EMG_{MIN}) / EMG_{MIN}) × $WEMG_{MAX}$
20 $W_{HRv} \leftarrow 1.0 - W_{EMG}$
21 $V \leftarrow R_{HRv} \times W_{HRv} + EVP \times W_{EMG}$
22 **return** (A, V)

Algorithm B. Rules for mapping the regressed physiological metrics into the AV dimensions in the manual approach.
Similarly to the grounded approach, the following algorithm was applied on top of the regressed physiological metrics to map them to arousal and valence. This method is a simple re-implementation of the current accepted good practices in physiological / affective computing.

Algorithm B: Manual Approach

Inputs:
Regressed arousal indicators
{ RSC, RHR}
Regressed valence indicators
{ $REMG_{ZYG}$, $REMG_{CORR}$}

Parameters:
None

Output:
Arousal prediction A
Valence prediction V

1 **if** $RSC = \emptyset$ **then**
2 $A \leftarrow RHR$
3 **elif** $RHR = \emptyset$ **then**
4 $A \leftarrow RSC$
5 **else**
6 $A \leftarrow (RSC + RHR) / 2$
7 **if** $REMG_{ZYG} = \emptyset$ **then**
8 $V \leftarrow REMG_{CORR}$
9 **elif** $REMG_{CORR} = \emptyset$ **then**
10 $V \leftarrow REMG_{ZYG}$
11 **else**
12 $V \leftarrow (REMG_{ZYG} + REMG_{CORR}) / 2$
13 **return** (A, V)

BREAKDOWN OF PLAYERS' EMOTIONAL STATES BY DEMOGRAPHIC AND GAMING CONDITIONS